University of Southampton

Faculty of Engineering and Physical Sciences

School of Engineering

Improving Objective Analysis of the Auditory Brainstem Response

by

Richard Michael McKearney

ORCID ID 0000-0001-7030-5617

Thesis for the degree of Doctor of Philosophy

May 2023

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences

School of Engineering

Doctor of Philosophy

Improving Objective Analysis of the Auditory Brainstem Response

by

Richard Michael McKearney

Auditory brainstem response (ABR) testing is a form of electrophysiological assessment used clinically to evaluate the auditory system. One of the main uses of ABR testing is in the evaluation of hearing thresholds in patients for whom behavioural hearing assessments are unreliable, e.g. newborns. Accurate interpretation of the ABR is important, as this will inform clinical decision making and potentially be used to prescribe hearing aid amplification. The overall aim of this research project was to explore methods for improving objective analysis of the ABR.

The first study in this thesis evaluated machine learning approaches for ABR detection. Using simulation, based on data recorded from participants, a range of machine learning algorithms were evaluated using nested k-fold cross-validation. The best algorithm, a stacked ensemble, was evaluated on previously unseen test set data. Using the bootstrap method to set the critical value for determining whether a response is present or absent, the stacked ensemble was able to achieve a high and stable level of specificity across ensemble sizes. Additionally, the detection rate of the stacked ensemble was statistically significantly better across all ensemble sizes, compared to the statistical detection methods evaluated. These results suggest that the proposed stacked ensemble algorithm may have the potential to assist clinicians in interpreting ABR waveforms, as well as in improving the performance of automated detection algorithms in ABR screening devices.

Due to the low signal-to-noise ratio of the ABR, detection of a response using visual inspection and statistical detection methods can be extremely challenging. Weighted averaging has been proposed as a method of maximising the signal-to-noise ratio in the averaged waveform. A second study aimed to further the understanding of weighted averaging, optimise the parameters of this technique, and quantify its effects on ABR detection using the Fmp statistical detection method. In this second study, the noise level estimation method was optimised, as was the parameter for the number of epochs in each block.

As well as being used for hearing threshold estimation, the ABR test may be used diagnostically in the functional assessment of the auditory brainstem pathway, e.g. for the detection of pathologies affecting the structures of this pathway. In a bid to reduce subjectivity in waveform interpretation, in study three of this thesis, several machine learning algorithms were compared in their ability to correctly predict ABR wave latencies. A convolutional recurrent neural network performed best, with 95.9% of predictions being within 0.1 milliseconds of the target label. Overall, this thesis provides three main approaches for improving objective analysis of the ABR. Further work is recommended to help translate this research into clinical practice.

Table of Contents

Table of Tables

Table of Figures

Table of Figures

List of Accompanying Materials

The ABR data used for the research in this have been described in previously published works (Lv, Simpson and Bell, 2007; Chesnaye *et al.*, 2018; Chesnaye, 2019), and were made available by Dr Michael Chesnaye in the University of Southampton Institutional Repository: doi:10.5258/SOTON/D0168.

Research Thesis: Declaration of Authorship

Print name: RICHARD MICHAEL MCKEARNEY

Title of thesis: Improving Objective Analysis of the Auditory Brainstem Response

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

McKearney, R. M. *et al*. (2022) 'Auditory Brainstem Response Detection Using Machine Learning: A Comparison With Statistical Detection Methods', *Ear & Hearing*, 43(3), pp. 949–960.

McKearney, R. M. et al. (2023) 'Optimising Weighted Averaging for Auditory Brainstem Response Detection', *Biomedical Signal Processing and Control*, 83, p. 104676. Available at: https://doi.org/10.1016/j.bspc.2023.104676.

Signature: ..... Date: 03/05/2023

Acknowledgements

I am extremely grateful to my supervisors Professor Steven Bell, and Professor David Simpson for their generous support, continual encouragement, and invaluable advice throughout my PhD studies. Thank you, Steve and David, for guiding me through the challenges that research presents and helping to turn them into learning opportunities. Thank you both also for the in-depth feedback and advice that you have provided on this thesis.

Thank you to Dr Michael Chesnaye, whose research has acted as an inspiration for the work presented in this thesis. Thank you, Michael, for the many thought-provoking discussions and for all your help and support. I am very grateful to Dr James Harte and Dr Sara Madsen for allowing use of their no-stimulus EEG database. I am also grateful to Debbie Cane who collected the ABR data used in this thesis, and to Dr Michael Chesnaye who processed this database, making it available via the University of Southampton Institutional Repository. I am grateful to Dr Michael Pont for allowing use of the diagnostic ABR database. Thank you to Dr Jaime Undurraga for your helpful advice regarding the effects of the analysis window length on the Fmp statistic.

Thank you also to Professor Thomas Blumensath and Dr Ben Lineton for the feedback provided on my progression review reports.

Thank you to the University of Southampton who funded my PhD studies. I would also like to gratefully acknowledge use of the IRIDIS High Performance Computing Facility, and the associated support services at the University of Southampton, in the completion of the research contained within this thesis.

I am grateful to my family and friends for their love and encouragement throughout my studies, in particular to my wonderful wife Julia and son Leo for their tremendous love and support.

Definitions and Abbreviations

AABR ..... Automated auditory brainstem response

ABR ..... Auditory brainstem response

AEP ..... Auditory evoked potential

AI ..... Artificial intelligence

ANSD ..... Auditory neuropathy spectrum disorder

AOAE ..... Automated otoacoustic emissions

ASSR ..... Auditory steady-state response

BSA ..... British society of audiology

CAEP ..... Cortical auditory evoked potential

CANS ..... Central auditory nervous system

CI ..... Confidence interval

CNN ..... Convolutional neural network

CNN-LSTM ..... Convolutional long short-term memory network

CNS ..... Central nervous system

CWT ..... Continuous wavelet transform

dB HL ..... Decibels hearing level

dB SL ..... Decibels sensation level

DCT ..... Discrete cosine transform

df ..... Degrees of freedom

DTW ..... Dynamic time warping

DWT ..... Discrete wavelet transform

ECG ..... Electrocardiogram

EEG ..... Electroencephalogram

FDB ..... Frequency domain bootstrap

FFT ..... Fast Fourier transform

FPR ..... False positive rate

HT2 ..... Hotelling's T^2

i.i.d. ..... Independent and identically distributed

LSTM ..... Long short-term memory

MAE ..... Mean absolute error

MLP ..... Multilayer perceptron

MMSE ..... Minimum mean square error

MRI ..... Magnetic resonance imaging

NHSP ..... Newborn hearing screening programme

NICU ..... Neonatal intensive care units

NOHL ..... Non-organic hearing loss

OCT ..... Optical coherence tomography

PCHI ..... Permanent childhood hearing impairment

PSD ..... Power spectral density

PTA ..... Pure tone audiometry

RMSE ..... Root-mean-square error

ROC AUC ..... Area under the receiver operating characteristic curve

SANR ..... Signal-amplitude-to-noise ratio

SD ..... Standard deviation

SE ..... Standard error

SNR ..... Signal-to-noise ratio

SVM ..... Support vector machine

T_s ..... Sampling interval

UI ..... User interface

UK ..... United Kingdom

VEP ..... Visual evoked potential

Chapter 1 Introduction

This thesis will present the findings of research undertaken with the aim of improving analysis of the auditory brainstem response (ABR). The ABR is an electrophysiological test which is used clinically to evaluate the auditory system. The test involves playing repeated auditory stimuli to the subject and recording the electrical response of the auditory brainstem via scalp surface electrodes. The electrical activity of the auditory brainstem is of a low signal-to-noise ratio and so each individual recording is dominated by background electrical activity. This background EEG is generated by the activity of neurones within the brain (Kirschstein and Köhling, 2009). By averaging together hundreds/thousands of recordings, the signal-to-noise ratio may be improved, allowing the ABR signal to be estimated (if it is present). The ABR may be used to objectively determine the hearing threshold of an individual. This is especially useful when audiologically assessing individuals who may be unable to undertake hearing tests requiring subjective input, e.g. newborns, hence the use of the ABR in the Newborn Hearing Screening Programme (NHSP) (British Society of Audiology, 2021), both as a screening test and as a diagnostic test for those newborns referred by the NHSP. The ABR is also used clinically in the neurological evaluation of the auditory brainstem pathway, helping to diagnose conditions affecting this pathway, such as vestibular schwannomas (tumours affecting the eighth cranial nerve). Chapter 2 of this thesis provides detailed background information regarding the anatomy and physiology which underpin the ABR, as well as an overview of the key clinical uses of this versatile test.

This thesis will explore how objective analysis of the ABR may be improved. The scope of this subject is very broad, and so this thesis will focus on three main research areas which shall be considered in turn. The first topic of this thesis focuses on improving ABR detection, specifically through the use of machine learning algorithms. Interpretation of the ABR is typically performed by trained clinicians who visually inspect the recorded waveforms. However, interpretation of waveforms, even amongst experienced clinicians, is known to be highly variable (Vidler and Parker, 2004). Statistical detection methods exist which may be used by clinicians to assist with ABR interpretation. Recent studies have examined how ABR detection may be improved using statistical methods (Chesnaye *et al.*, 2018; Chesnaye, 2019). A background of ABR detection using statistical detection methods is presented in Chapter 3. There has also been interest in applying machine learning techniques to this clinical challenge (Acir, Erkan and Bahtiyar, 2013; McKearney and MacKinnon, 2019). Given the successes of recent studies using machine learning algorithms in a variety of related signal processing studies (Hannun *et al.*, 2019; Medvedev, Agoureeva and Murro, 2019), there is potential that machine learning algorithms may be applied to this field in order to further improve detection performance. The topic of ABR detection using machine

learning is explored in Chapter 4, where it will be shown that machine learning algorithms have the potential to outperform the prominent, more conventional, statistical ABR detection methods. Improved ABR detection performance could lead to improved performance of newborn hearing screening programmes, e.g. by reducing the number of cases where a hearing loss fails to be correctly detected and/or reducing the number of cases where an individual with normal hearing is incorrectly identified as having a hearing loss. Improving detection of hearing loss at an early age allows the early provision of audiological habilitation, which has been shown to lead to improved receptive and expressive language development (Yoshinaga-Itano, Coulter and Thomson, 2001; Pimperton and Kennedy, 2012).

The second topic explored in this thesis is how weighted averaging may be optimised to improve ABR detection (Chapter 5). Detection of the ABR signal relies on averaging together numerous (hundreds/thousands) individual recordings. Unweighted averaging is suboptimal if the background noise in the recording is non-stationary (Hoke *et al.*, 1984). Weighted averaging has therefore been proposed in order to provide a better estimate of the ABR signal within the averaged waveform (Elberling and Wahlgreen, 1985). However, there has been relatively little work in the field on how to optimise the weighted averaging technique, e.g. the noise level estimation method and the optimal block size to use. The second topic in this thesis therefore focuses on how weighted averaging may be optimised in order to improve ABR detection. Incremental gains in detection performance may lead to a large population-level benefit, especially given the widespread use of the ABR as part of newborn hearing screening programmes around the world. Chapter 5 will show how the parameters of weighted averaging may be optimised in order to improve detection of the ABR. This chapter will also show how weighted averaging performs in combination with statistical detection methods (the Fmp), as well as how changes to the false positive rate may be overcome by applying the bootstrap method to the weighted test statistic (Lv, Simpson and Bell, 2007; Chesnaye *et al.*, 2018).

The third and final topic of this thesis is on improving analysis of the diagnostic ABR using machine learning (Chapter 6). Whilst the first two topics of this thesis have focused on improving detection of the ABR in order to better estimate hearing thresholds, this third topic centres around a different clinical use for the ABR: diagnostic evaluation of the auditory brainstem pathway. When a sound is played at suprathreshold levels, the full morphology of the ABR waveform and all of its component waves becomes evident (Jewett, Romano and Williston, 1970). The morphology of the ABR waveform represents the function of the structures that make up the auditory brainstem pathway, and so abnormalities in the waveform may be used to detect pathology affecting this pathway, e.g. tumours of the vestibulocochlear nerve. This third study applies machine learning algorithms to the analysis of the diagnostic ABR. Specifically, the proposed technique aims to label

the key waves of the ABR waveform, estimating their latencies, which is useful for clinical decision making (British Society of Audiology, 2019b). Chapter 6 will show how machine learning algorithms can successfully label the key waves of the ABR waveform, as well as provide a measure of confidence to assist in interpretation of the results.

The main aims of this thesis were therefore as follows:

1.1.1 ABR Detection using Machine Learning

1. Develop a suitable database of 'response present' and 'response absent' data by which to train and evaluate machine learning algorithms.
2. To train a machine learning algorithm to be able to determine whether EEG data contains an ABR or not.
3. Compare the performance of the machine learning algorithm with that of prominent statistical ABR detection methods.

1.1.2 Automated ABR Detection and Weighted Averaging

1. To optimise weighted averaging by identifying the value of the epochs-per-block parameter that reduces noise within the averaged waveform and improves ABR detection the most.
2. Compare methods of estimating the variance of the noise level within each block, to further optimise weighted averaging.
3. Investigate the effects of weighted averaging on the Fmp statistical ABR detection method.

1.1.3 Automated Analysis of the Diagnostic ABR using Machine Learning

1. To propose, train, and evaluate automated machine learning algorithms which are able to label waves I, III and V of the diagnostic ABR. Multiple state-of-the-art algorithms should be evaluated to select the best approach. The automated algorithm should also provide a confidence measure to help clinicians interpret the latency values provided. The aim is not to present a final model, ready for clinical implementation, but rather to identify promising algorithms which may then be evaluated on larger datasets reflective of the intended clinical population.

Whilst a brief introduction has been provided here, a more detailed introduction, including a literature review for each topic, will be provided in the relevant chapter.

1.2 Research Hypotheses

The main hypotheses which will be evaluated in this thesis are outlined below.

1.2.1 ABR Detection using Machine Learning

1. Trained machine learning algorithms can provide a more effective method of detecting the ABR compared to prominent statistical detection methods, specifically with regard to sensitivity and specificity.

1.2.2 Automated ABR Detection and Weighted Averaging

1. ABR detection may be improved by more accurately estimating the variance of the background noise, using the 'VAR Whole Block' method, compared to the 'VAR MP' method.

1.2.3 Automated Analysis of the Diagnostic ABR using Machine Learning

1. Machine learning algorithms may be trained to accurately estimate the latency of ABR waves I, III, and V, performing better than a baseline estimator.
2. Confidence predictions for wave latency estimates, provided by machine learning algorithms, will be able to reflect those provided by a human clinician as measured by their correlation.

1.3 Research Significance

The ABR forms a critical component of the Newborn Hearing Screening Programme (Public Health England, 2020). The performance of the Newborn Hearing Screening Programmes is therefore intricately linked to the performance of the ABR detection algorithm used (as well as the performance of the otoacoustic emissions test used). Approximately 660,000 babies are born in the England per annum (Wood, Sutton and Davis, 2015). The coverage (i.e. uptake) of the NHSP is 98.95% (Wood, Sutton and Davis, 2015). The referral rate of the NHSP in England is between 2–3% (Wood, Sutton and Davis, 2015). This means that approximately 16,000 babies per annum (minus the number of babies referred direct for follow-up assessment without screening, e.g. due to microtia) will have an automated ABR (AABR) test and not pass it in one or both ears, leading to a referral from the NHSP for follow-up audiological assessment. A number of other babies will have had the AABR and passed, and therefore not be referred on. Babies who are referred by the NHSP will typically have diagnostic ABR testing to objectively establish their hearing thresholds

(British Society of Audiology, 2021). In England alone, the use of the ABR in the newborn population is widespread. This is reflected globally, to various extents, by other national hearing screening programmes (New Zealand Ministry of Health, 2016; Linnebjerg, Hansen and Møller, 2017). Improvement in the performance of the ABR test, which was the main aim of this thesis, therefore has the potential to positively influence the performance of these screening programmes. This may be through improved screening programme sensitivity, where a greater number of cases of newborns with hearing loss are correctly identified which would otherwise have been missed. This would allow audiological support for their hearing loss to be initiated earlier, leading to improved outcomes (Pimperton and Kennedy, 2012). Improved test performance may also be reflected by an improved screening programme specificity, whereby fewer newborns with normal hearing are referred on for further diagnostic testing. Increasing screening programme specificity would therefore prevent unnecessary stress and anxiety for parents/carers by reducing the number of newborns with normal hearing who are referred on for further testing, as well as saving administrative and clinical time. Improving ABR detection has the potential to decrease the time required to complete testing (Chesnaye *et al.*, 2018). This is helpful as ABR testing is typically performed on newborns when they are sleeping; testing may be stopped prematurely if the newborn wakes up, potentially necessitating a further appointment to complete testing (British Society of Audiology, 2021). The ABR is also used for hearing threshold estimation in older children and adults for whom subjective hearing evaluation may potentially be unreliable, e.g. some individuals with a learning disability. Improving the performance of ABR detection algorithms therefore has the potential to assist clinicians in providing more effective audiological care in a variety of clinical situations. Benefits in ABR detection also have the potential to be transferred to other evoked potential modalities, e.g. the visual evoked potential, and the somatosensory evoked potential (Walsh, Kane and Butler, 2005).

The diagnostic ABR is used in a variety of clinical situations to help diagnose pathologies affecting the structures which contribute to the auditory brainstem pathway. By using machine learning algorithms to improve the estimation of the latency of the key waves of the ABR, it is anticipated that clinicians may be supported in interpreting the ABR waveform, improving the accuracy and consistency of interpretation.

1.4 Original Contributions

1.4.1 ABR Detection using Machine Learning

The work presented in Chapter 4 focuses on how machine learning algorithms may be effectively trained to detect the ABR. Previous studies in this field have suffered from the limitation that the

true labels of the data (whether a response is present or absent) are unknown (Alpsan, 1991; Bradley and Wilson, 2005; McKearney and MacKinnon, 2019). Unlike previous studies in the field of ABR detection using machine learning, this work uses simulated ABR data in order to overcome this limitation. Due to differences in methodology, datasets, and outcome measures used, it is challenging to compare the performance of different machine learning algorithms developed to detect the ABR (McKearney *et al.*, 2022). An original contribution of this work is that the performance of the presented machine learning algorithm is compared to that of prominent established statistical detection methods. The presented algorithm is the first one to have been objectively demonstrated to exceed the performance of statistical ABR detection methods. The comparison with statistical ABR detection methods means that, even if future studies use different datasets and algorithms, a form of relative performance comparison will be able to be drawn via a comparison between the proposed algorithm and statistical detection methods which are relatively straightforward to implement. A further original contribution of this work is that it represents the first reported application of the previously developed ABR bootstrap technique (Lv, Simpson and Bell, 2007) to an ABR machine learning detection algorithm, showing it to be effective at controlling the false positive rate, obviating the need to define the detection criterion using a separate set of data.

1.4.2 Automated ABR Detection and Weighted Averaging

The work presented in Chapter 5 is focused on improving weighted averaging in terms of ABR detection. In general terms, weighted averaging involves estimating the noise level within one or more recording epochs and then weighting that block of recording epochs inversely proportional to the estimated noise level. This upweights the information provided by epochs with low noise, and downweights the information from epochs containing lots of background noise (Elberling and Wahlgreen, 1985). Whilst previous studies have already provided guidance on the optimal block size for weighting averaging (Elberling and Wahlgreen, 1985; Don and Elberling, 1994; Riedel, Granzow and Kollmeier, 2001), the work presented in this thesis provides detailed additional evidence, optimising the block size for weighted averaging both in terms of residual noise levels and for ABR detection using the Fmp test statistic. Elberling and Wahlgreen (1985) provide some evidence for the effects of weighted averaging on the Fmp statistic in the form of a few individual recordings. The presented study, however, through the use of subject recorded background EEG data and simulated ‘response present’ data, provides an in-depth analysis of the effects of weighted averaging on the Fmp test statistic. The experiment presented in this thesis measured the mean null condition Fmp statistic to have a value below the expected value of around one. After discussion with colleagues in the field of evoked potential research (J. Undurraga, personal

communication, 2022), this finding was shown to be due to the finite length of the analysis window. Whilst this potential effect was theorised by Elberling and Don in 1984, the experiment published in this thesis describes the effect that this has on ABR detection when combined with weighted averaging. The presented work also shows how the characteristics of the data affecting statistical test performance when coupled with weighted averaging, may be controlled for by using the bootstrap method (Lv, Simpson and Bell, 2007), leading to improved ABR detection with a controlled specificity level.

Within the weighted averaging procedure, the noise level in groups of recording epochs may be estimated using either a single point (Elberling and Wahlgreen, 1985) or multiple points method ('VAR MP'), whereby the average of the variance of samples down one (single point) or multiple (multiple point) columns within a block of epochs is taken. A further original contribution is through the comparison of the 'VAR MP' method with a method whereby the variance of all the samples within the block are used as the noise level estimate ('VAR Whole Block'), with the 'VAR Whole Block' method being shown to be superior for the dataset used in the study in terms of reducing residual noise and improving ABR detection using the Fmp statistic (McKearney et al., 2023).

1.4.3 Automated Analysis of the Diagnostic ABR using Machine Learning

The work in Chapter 6 presents the findings of a study using machine learning to estimate the latency of waves I, III, and V of the ABR. The presented work represents the first instance in the literature of convolutional layers being used for analysis of the diagnostic ABR. The best algorithm presented (a convolutional recurrent neural network) performs better than the state-of-the-art algorithms described in the literature (Chen *et al.*, 2021). However, it is acknowledged that the methodological differences between the studies, as well as the data used, make drawing direct comparisons difficult. An additional original contribution is that, unlike most other studies in the field of using machine learning to label the waves of the ABR, the present study provides an algorithm which predicts a confidence measure for how likely it is that an accurate wave latency prediction can be made. Whilst the rule-based algorithm provided by Boston (1989) provides a confidence measure for the wave V latency prediction, the proposed method uses a neural network to make this confidence level prediction. Providing a confidence level estimate may help clinicians to be able to better interpret the wave latency predictions, by paying greater heed to the predictions where there is high confidence and relying to a greater extent on their own visual interpretation of the waveform when a low confidence prediction is provided.

1.5 A Note on the Format of this Thesis

The work in this thesis centres around three main studies presented in Chapters 4, 5, and 6. As well as the Introduction and Conclusions chapters, which bring the overall work in the three mains studies together, further study-specific literature reviews, discussion and conclusions are presented in the relevant chapters. This format was selected in order to present the relevant information in a coherent and accessible format.

1.6 Publications and Presentations

1.6.1 Published Articles

McKearney, R. M., Bell, S. L., Chesnaye, M. A., and Simpson, D. M. (2022) 'Auditory brainstem Response Detection Using Machine Learning: A Comparison with Statistical Detection Methods', *Ear & Hearing*, 43(3), pp. 949–960.

McKearney, R. M., Bell, S. L., Chesnaye, M. A., and Simpson, D. M. (2023) 'Optimising Weighted Averaging for Auditory Brainstem Response Detection'. *Biomedical Signal Processing and Control*, 83, p. 104676. Available at: https://doi.org/10.1016/j.bspc.2023.104676.

1.6.2 Planned Article Submissions

McKearney, R. M., Bell, S. L., and Simpson, D. M. 'Objective Analysis of the Diagnostic Auditory Brainstem Response using Machine Learning'.

1.6.3 Conference Presentations

McKearney, R. M., Bell, S. L., Chesnaye, M. A., and Simpson, D. M. (2021) 'Detecting the ABR using Machine Learning'. XXVII Symposium International Evoked Response Audiometry Study Group. Online (virtual conference). *Recorded oral presentation.*

McKearney, R. M., Bell, S. L., and Simpson, D. M. (2022) 'Analysing the diagnostic auditory brainstem response using machine learning'. UK 'Ear and Hear' Meeting, The UK Acoustics Network. Southampton, UK. *Poster presentation.*

Chapter 2 The Auditory Brainstem Response

2.1 Background and Physiology

2.1.1 The Human Auditory System

This chapter provides an overview of the core topic of this thesis—the auditory brainstem response (ABR). In order to describe the ABR in a meaningful context it is useful to first briefly consider the auditory system as a whole.

The human auditory system is a remarkable apparatus which operates over a wide dynamic range and is capable of detecting and interpreting tiny fluctuations in air pressure: for example, for 1kHz tones, a healthy human ear can detect sounds pressures at a level of 47 µPa (Gelfand, 2009). Hearing is useful not only in detecting environmental sounds but is also crucial for speech perception. The peripheral auditory system comprises the structures of the outer, middle, and inner ear up to where the cochlear nerve ends where it connects at the brainstem (Figure 2-1A). The role of the peripheral auditory system is to convert fluctuations in air pressure into electrochemical signals which can subsequently be interpreted by the brain. The anatomy of the auditory system comprises a variety of specialised components which enable this process to occur (Figure 2-1A). The outer ear consists of the pinna and ear canal, which direct sound waves to the middle ear via the tympanic membrane. The middle ear acts as a transformer, allowing the sound waves to be transferred from the air around us into the fluid-filled cochlea via the oval window without being severely attenuated (Wilson, 1987). This is achieved by the lever action of the three small bones of the middle ear (ossicles) as well as the advantageous area ratio between the tympanic membrane and round window (Wilson, 1987). Vibrations of the tympanic membrane therefore lead to displacement of the oval window and subsequently displacement of the perilymph fluid within the cochlea. The transduction of sound (mechanical energy) into electrochemical energy takes place in the cochlea (Naftalin, 1981). This process is necessary in order for the sound to be transmitted via neurones to the brain. Inner hair cells are the site of this transduction, with inner hair cells being displaced by vibrations of the fluid in the cochlea leading to their depolarisation (Zwislocki, 1980; Kiang *et al.*, 1986). This cascades into the release of neurotransmitters which cause the cochlear nerve to fire an action potential.

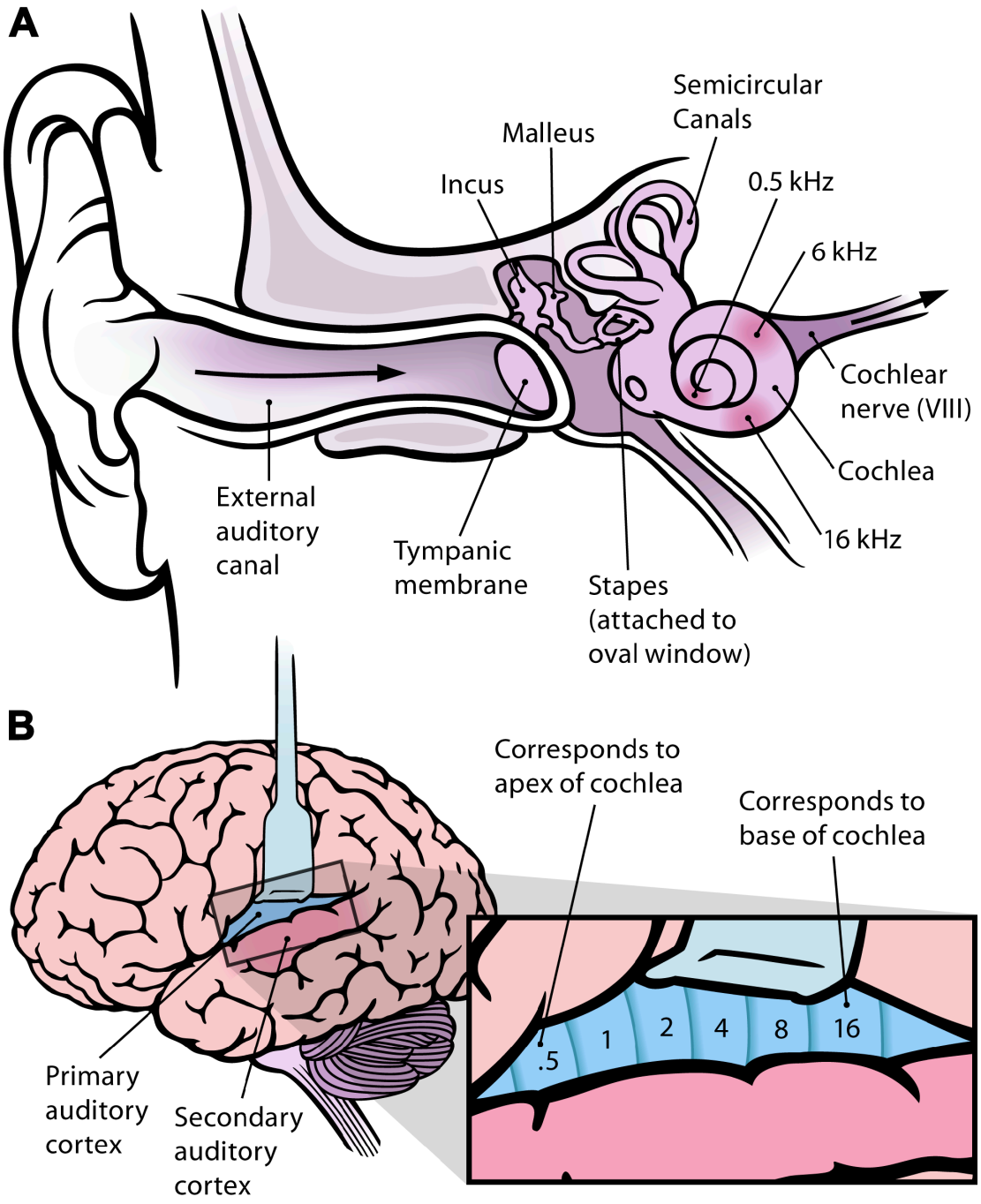

Figure 2-1 The human auditory system. Figure **A** shows the anatomical structures of the peripheral auditory system: the outer, middle and inner ear. Figure **B** shows the auditory cortex. This figure ('Frequency Coding in the Human Ear and Cortex') is reproduced, with no changes made, from Chittka, L. and Brockmann, A. (2005) 'Perception Space—The Final Frontier', PLoS Biology. Public Library of Science, 3(4), p. e137. Available at: https://doi.org/10.1371/journal.pbio.0030137. This figure is reproduced in accordance with the terms of the CC BY 2.5 license under which the image was published.

It is the role of the central auditory nervous system (CANS) to process and interpret the information received (Figure 2-1A and Figure 2-2). It is in the CANS where the complex tasks of sound recognition and sound localisation take place (Staecker and Thompson, 2013).

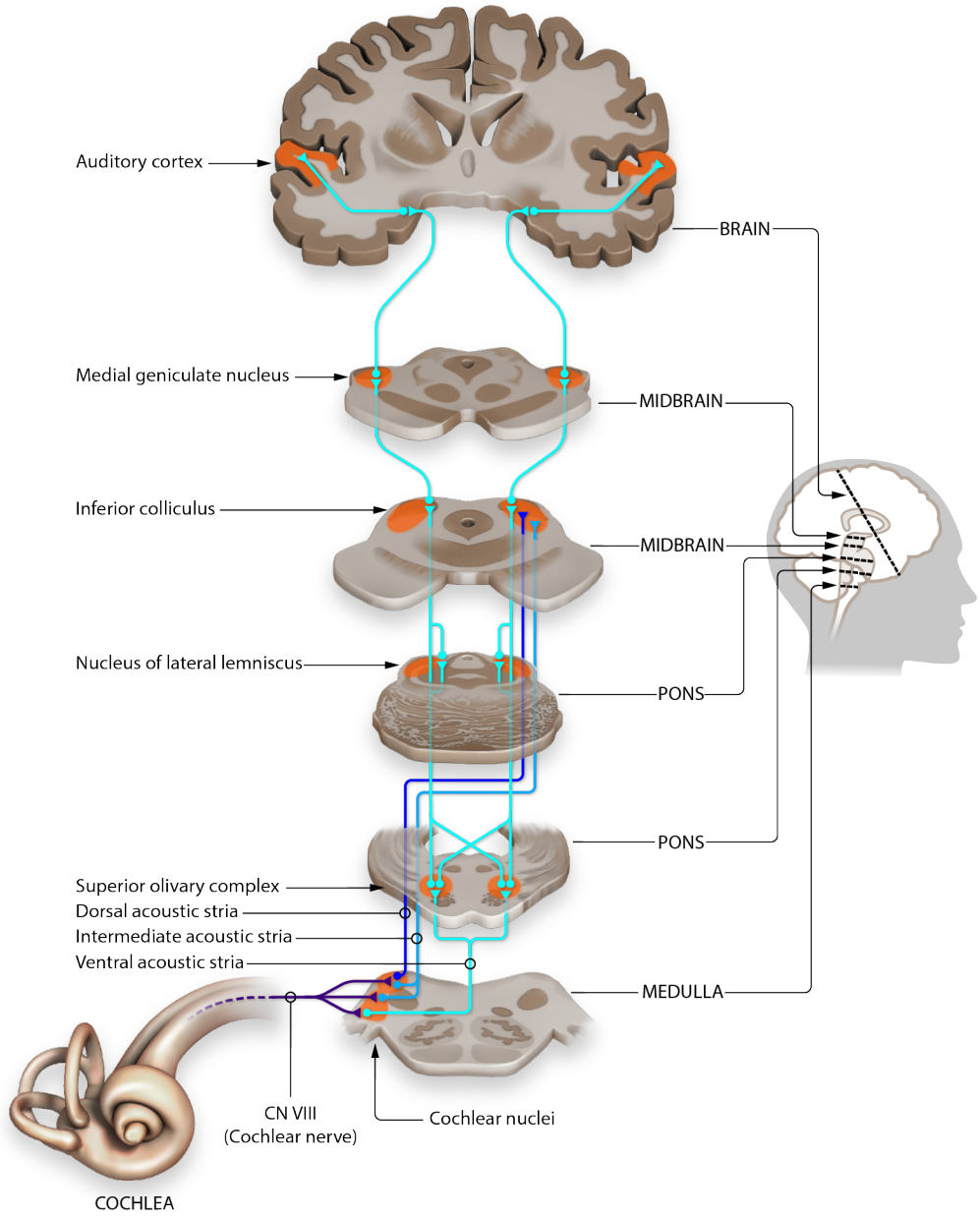

Figure 2-2 The auditory nervous system. This figure is reproduced, from Peelle, J. E. (2016) Human Auditory Pathway, available at: https://osf.io/u2gxc/. This figure is reproduced under the terms of the CC BY 4.0 license, with the image changed to crop out the rest of the figure.

2.1.2 Auditory Evoked Potentials

Auditory Evoked Potentials (AEPs) are measurements of auditory system activity in response to a sound (auditory stimulus). When an auditory stimulus is delivered to the ear, a cascade of

electrochemical activity occurs, converting a sound from a mechanical wave into a format interpretable by our brain. Electrophysiological recording equipment can be used to measure these electrochemical events. The traditional hearing test (audiometry) where a patient is asked to press a button in response to a sound is a form of behavioural test as it relies on the subjective behavioural response of the participant. Unlike behavioural tests, evoked potentials do not require the patient's subjective input. They are therefore considered to be an objective form of hearing assessment. Evoked responses are typically recorded as voltage amplitudes fluctuating over time and can be used to make inferences regarding the status of an individual's auditory system (Stapells, 2000; Gorga *et al.*, 2006). This makes AEPs useful clinically. Evoked potentials are recorded as a potential difference between two recording electrodes. Clinically, AEPs such as the ABR are measured as far-field recordings, i.e. they are not measured directly from neurological generators close to the electrode, but are instead recorded at a distance from the source of the electrical activity by electrodes placed in standard positions on the scalp (Figure 2-3).

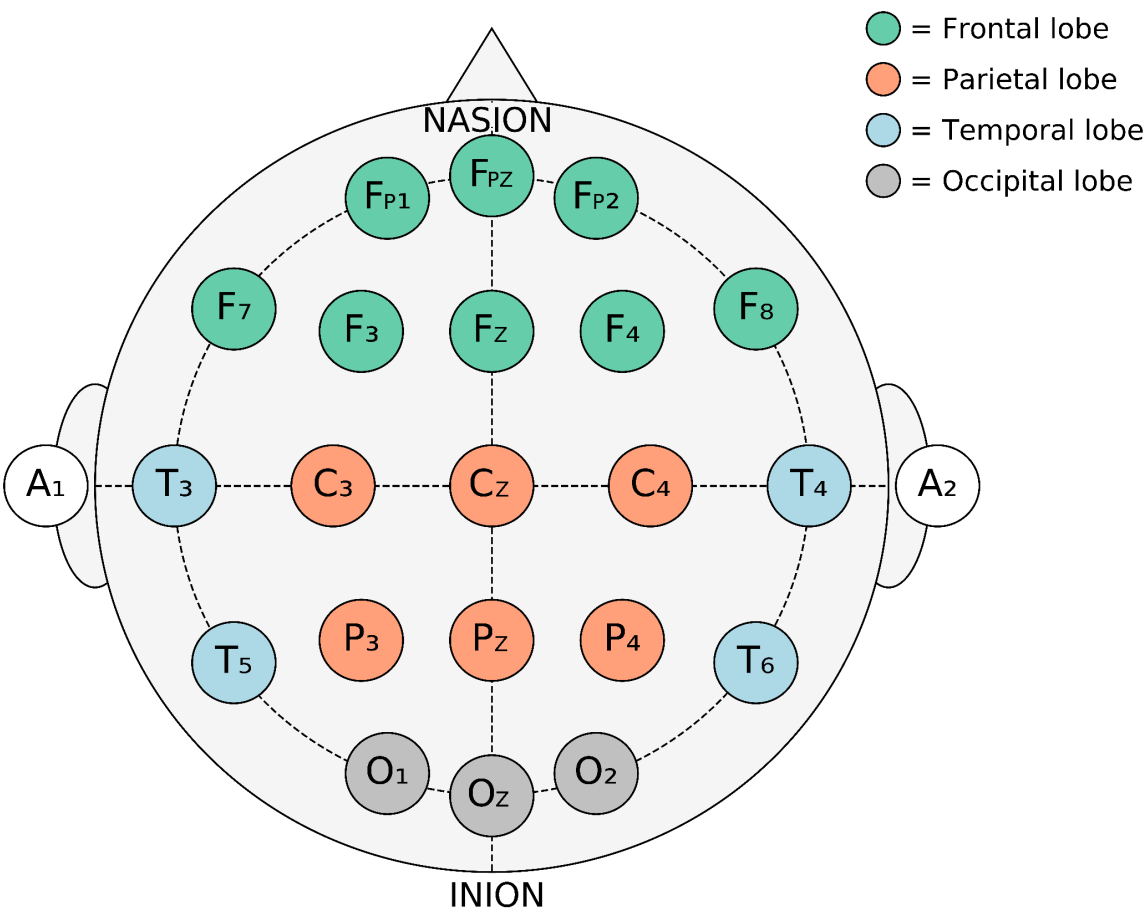

Figure 2-3 International 10–20 system of electrode placement as described by Klem *et al.* (1999). Figure redrawn based on a figure in Kim, J. H., Kim, C. M. and Yim, M. S. (2020) 'An Investigation of Insider Threat Mitigation Based on EEG Signal Classification', Sensors 2020, Vol. 20, Page 6365. Multidisciplinary Digital Publishing Institute, 20(21), p. 6365. Available at: https://doi.org/10.3390/s20216365. This work was published by MDPI under a CC BY 4.0 license. This figure is redrawn, with changes made (added text/extra electrode positions/different colour scheme), under the terms of this license.

Using scalp recording electrodes makes the test non-invasive, but also means that the recorded voltages are smaller, especially in relation to the background electrical noise from ongoing EEG and muscular activity (Stegeman *et al.*, 1997). As an example the amplitude of the ABR signal can be up to ~500 nV (Hall, 2007), with background noise derived from unrelated neuronal activity, myogenic (muscular) activity and electrical interference having an amplitude upward of 15 µV.

There are several different AEPs whose nomenclature is typically informed by their physiological site of origin and therefore latency, with latency referring to how long after the auditory stimulus the response occurs (Figure 2-2). Responses which occur peripherally (i.e. earlier on in the auditory pathway) such as the cochlear microphonic, an alternating current potential reflecting

outer hair cell function (Hall, 2007), will have shorter latencies compared to responses of a central origin, e.g. the cortical auditory evoked potential (CAEP). Another significant distinguishing factor which classifies AEPs is whether they are transient or steady-state responses. Transient evoked responses, such as the ABR, are transient fluctuations which occur after the onset of each single stimulus. Steady-state responses, such as the auditory steady-state response (ASSR), are continuous and elicited by a rapidly periodically repeating stimulus (Galambos, Makeig and Talmachoff, 1981). Whilst there are numerous AEPs, the research in this thesis will focus on the ABR due to its widespread clinical adoption and prominent use in the UK Newborn Hearing Screening Programme (NHSP) (British Society of Audiology, 2019c). However, due to their common underlying principles, many of the findings will likely generalise not just across all transient AEPs but also to transient evoked potentials of other modalities, e.g. visual and somatosensory evoked potentials.

2.1.3 The Auditory Brainstem Response

The ABR first appeared in the literature in a study by Sohmer & Feinmesser (1967), however, the recording was initially interpreted as being a cochlear action potential. The response was first correctly ascribed as auditory brainstem activity by Jewett *et al*. (1970) who related the latency of the response to its physiological source of origin (Atcherson, 2012). The ABR waveform comprises of a series of peaks and troughs which give the response a characteristic appearance (morphology).

To record the ABR, the electrode configuration is chosen to reflect the orientation of the electrical dipole created by the flow of current during the evoked potential, thus maximising the amplitude of the recorded response (Stegeman *et al.*, 1997). An active ('positive') recording electrode is typically sited at Fz/Cz (high forehead/on top of the head), with the reference ('negative') electrode positioned on either the ipsilateral mastoid bone (British Society of Audiology, 2019c), or the nape of the neck (King and Sininger, 1992) (Figure 2-3). The largest component of the ABR waveform (wave V) arises from neurological generators with a vertical alignment; the typically vertically aligned electrode montage capitalises on this (Atcherson, 2012). The neurological generators of the ABR extend from the distal part of the auditory nerve, possibly up to the thalamic level (the medial geniculate body) (Møller *et al*, 1981; Hashimoto *et al*, 1981 in Hall, 2007) (Figure 2-2). The generators of the later ABR components are complex, poorly defined, and a subject of ongoing debate (Hall, 2007).

2.1.4 Clinical Use of the ABR

Of the AEPs, the ABR is one of the most commonly used tests in clinical practice (Atcherson, 2012), owing largely to the ability of the test to provide reliable estimates of an individual's hearing thresholds (Stapells, 2000; Gorga *et al.*, 2006). The groups of patients for whom objective ABR assessment is considered most useful typically reflects those patient groups who may not be expected to be able to readily engage with behavioural hearing assessments: newborn babies, infants, some adults with learning disabilities, some individuals with cognitive impairment, e.g. dementia, and some individuals with a non-organic hearing loss (NOHL) (British Society of Audiology, 2019c, 2019b). NOHL is commonly defined as 'responses to hearing tests indicating deficits that cannot be explained by known pathology' (Austen and Lynch, 2009). The caveat 'some' has been applied to some of the patient populations in the aforementioned list as the test strategy employed will very much be decided on an individual basis. For example, many adults with learning disabilities will be able to readily perform behavioural hearing assessments, producing reliable and accurate results. As well as being used for hearing threshold estimation, the ABR can provide insight into the neurological state of the auditory system by providing qualitative information regarding the function of the auditory nerve and auditory brainstem. These two functional use cases of the ABR shall be considered in turn.

2.1.4.1 Hearing Screening

In the UK (Public Health England, 2020), as well as many other countries (New Zealand Ministry of Health, 2016; Linnebjerg, Hansen and Møller, 2017), the ABR has a key role in the early identification of hearing loss in newborns. Permanent Childhood Hearing Impairment (PCHI) may either arise congenitally (e.g. genetically), or be acquired (e.g. secondary to congenital infection, hypoxia, ototoxicity, bacterial meningitis etc.) (Billings and Kenna, 1999). Early identification of hearing loss is important as it allows management strategies such as hearing aids or cochlear implantation to be initiated early in the baby's life (Pimperton and Kennedy, 2012). Providing babies access to sound is critical to their effective speech development and is associated with numerous positive health and educational outcomes throughout life (Yoshinaga-Itano *et al.*, 2001; Kennedy *et al*., 2006; Fulcher *et al*., 2012). Early intervention is critical as it allows a critical window in the brain's development to be harnessed, during a time when neural plasticity is high (Cardon, Campbell and Sharma, 2012). There is strong evidence for the need for early detection and treatment of hearing loss in babies and is the rationale for newborn hearing screening programmes. The UK Newborn Hearing Screening Programme (NHSP) relies on two forms of objective hearing assessment techniques, namely automated versions of the otoacoustic emissions (OAE) test and the ABR (Figure 2-4). Newborns who do not attain a clear response (CR)

in both ears for all three steps of the screening process are referred urgently to audiology for definitive assessment of their hearing (Public Health England, 2020). In this appointment the audiologist will typically perform ABR measurements to establish the hearing thresholds of the referred newborn (British Society of Audiology, 2021).

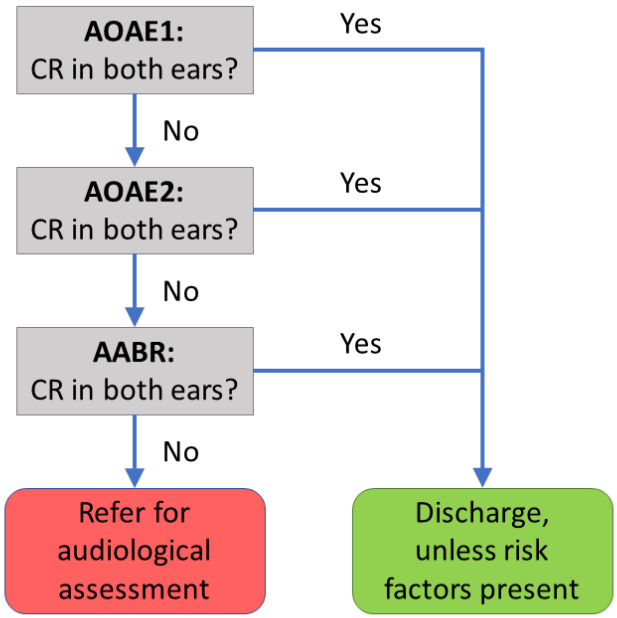

Figure 2-4 Summary of the Newborn Hearing Screening Programme pathway for well babies, for babies with no contraindications for screening. Please see Public Health England, 2020), for full details of the pathway. Babies with certain risk factors are at a higher risk of PCHI. Such risk factors include syndromes associated with hearing loss, e.g. Down syndrome, craniofacial abnormalities, e.g. cleft palate, congenital toxoplasmosis or rubella infection, amongst others (Public Health England, 2019). Babies with such risk factors should be referred for targeted audiology follow-up assessment (behavioural audiological assessment at around 8 months of age), even if AOAE1, AOAE2, and AABR provide a clear response (Public Health England, 2019). Note that there is a separate for babies in neonatal intensive care units (NICU) **AOAE** = Automated Otoacoustic Emissions; **AABR** = Automated Auditory Brainstem Response; **CR** = Clear Response. Contains public sector information licensed under the Open Government Licence v3.0. Figure redrawn with permission, with changes made, based on a figure produced by Public Health England (2020). Available at: https://www.gov.uk/government/publications/newborn-hearing-screening-care-pathways/newborn-hearing-screening-programme-nhsp-care-pathways-for-well-babies.

2.1.4.2 Hearing Threshold Estimation

Where behavioural hearing thresholds are unobtainable, e.g. in newborns, or otherwise not reliable, the ABR provides an objective method of assessing an individual's hearing thresholds. An example of the ABR recorded across a range of stimulus levels is shown in Figure 2-5. During the ABR test, an auditory stimulus is played repeatedly into the ear of the patient whilst the EEG is

being concurrently recorded. A 'loud' sound stimulus played above the individual's hearing threshold is expected to elicit a response of large amplitude (Hecox and Galambos, 1974). As the auditory stimulus level is reduced, the ABR amplitude is known to decrease, whilst the latency of the response increases (Hecox and Galambos, 1974) (Figure 2-5). As the auditory stimulus level is lowered further, there will be a point when the ABR is no longer detectible (Hall, 2007). The point at which the ABR is *only just* detectible is considered to be the ABR threshold. The clinical utility of this arises from the knowledge that the ABR threshold typically coincides closely with an individual's behavioural hearing threshold (Stapells, 2000; Gorga *et al.*, 2006). This means that the estimated thresholds obtained can be used to: inform diagnostic decision-making, programme hearing aids (using a correction factor) (Bagatto *et al.*, 2005), and to help provide guidance on effective communication tactics to parents/guardians/carers.

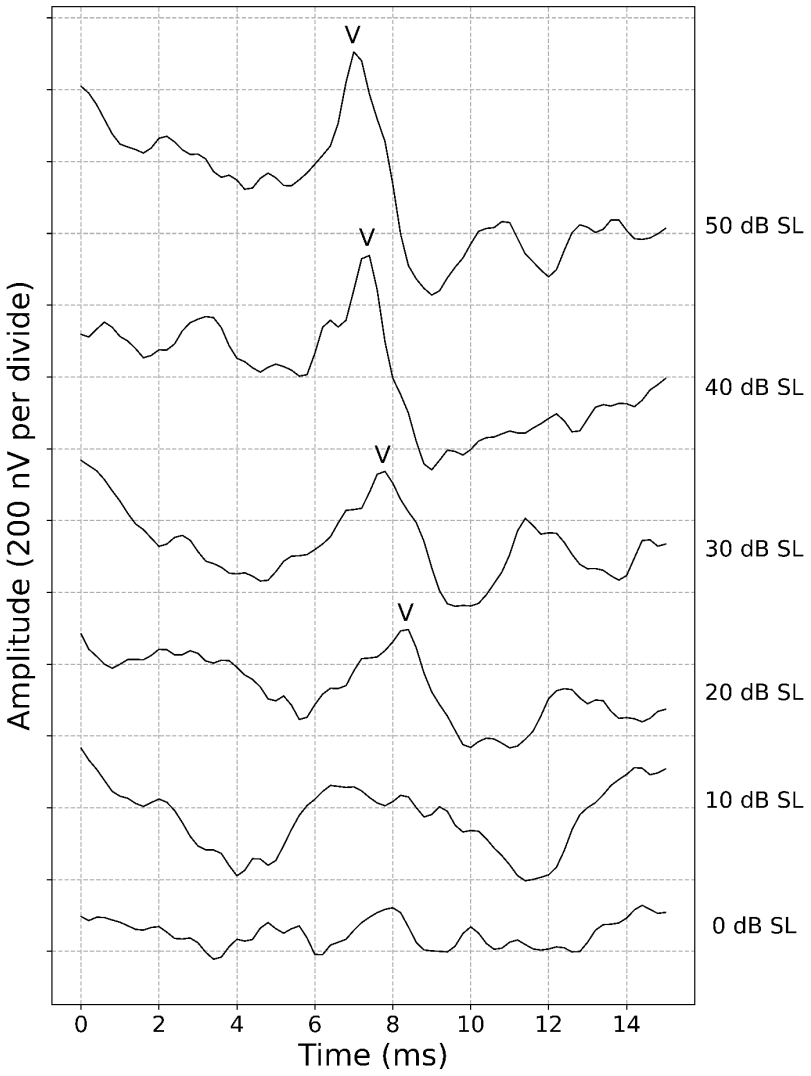

Figure 2-5 The auditory brainstem response from one adult with normal hearing. ABR recordings are shown across a range of stimulus levels from 0 to 50 dB SL (Sensation Level—relative to the individual's audiogram threshold). Where deemed present,

wave V has been labelled. Note the increasing wave V latency and reduced amplitude with decreasing stimulus level.

The ABR is clearly a very useful testing in screening the hearing of newborn babies. In addition to this vital role in detecting congenital or early acquired causes, the ABR may also be used at a later stage in an infant's life to detect hearing loss which may be acquired after this period, e.g. following exposure to ototoxic medication (medication with the side-effect of harming the ear) or bacterial meningitis.

2.1.4.3 Neurological Assessment

As well as being used for hearing threshold estimation, the ABR can be used to evaluate the neurological functioning of part of the auditory nervous system (the neurological structures which contribute to the generation of the ABR—Figure 2-2). The ABR when used for this purpose may be referred to the neurodiagnostic or neurological ABR (Figure 2-6). In particular the ABR has been widely used for evaluating the integrity of the auditory nerve (Selters and Brackmann, 1977; Schmidt *et al.*, 2001). Conditions affecting the integrity of the auditory nerve can lead to disruption in the ABR morphology including increased latencies of ABR peaks. For example, acoustic neuromas (benign tumours affecting the vestibulocochlear nerve) may impair the function of the auditory nerve, leading to delayed nerve propagation and an increase in the latency of waves III and V (generated by structures proximal to the disruption). The latency of

wave I (generated by the distal portion of the auditory nerve), would be expected to display a normal latency as it is distal to the site of lesion.

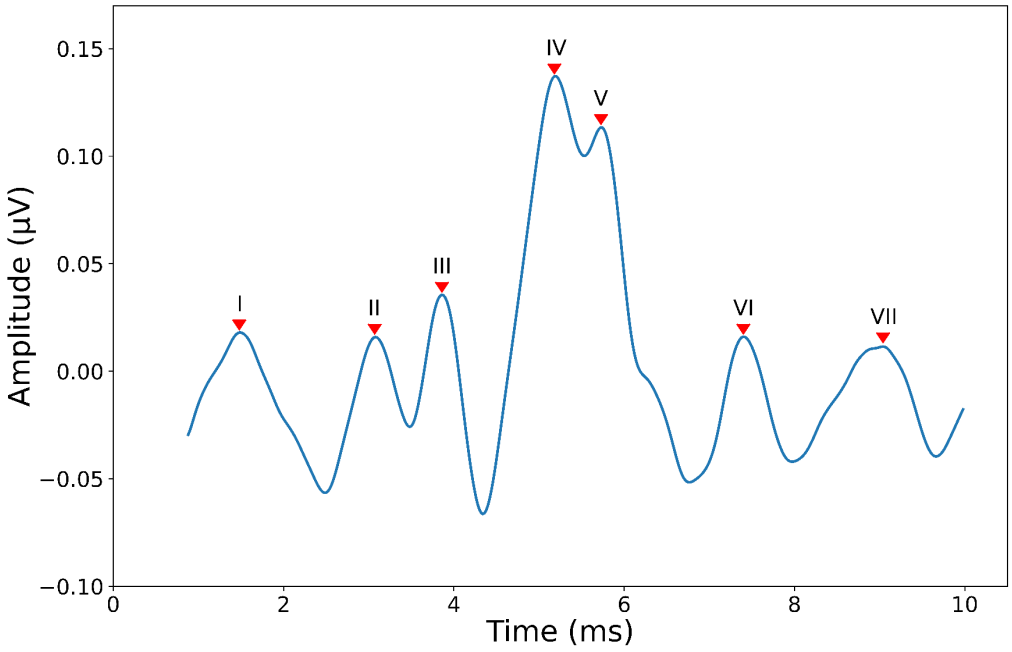

Figure 2-6 Neurodiagnostic ABR. This Figure shows an example ABR waveform recorded using a neurodiagnostic protocol from the database recorded by Sundaramoorthy *et al.* (2000). Waves I–VII are labelled using the Roman numeral convention established by Jewett *et al*. (1970).

In the case of an acoustic neuroma, the latency of wave I (generated by the distal portion of the auditory nerve—distal to the lesion), would be expected to display a normal latency. Reliance on the ABR for retrocochlear (proximal to the cochlea, i.e. auditory nerve and central auditory system) disease detection has been diminished by the availability of high-resolution magnetic resonance imaging (MRI). However, the need for the ABR has not been entirely superseded as it still acts as an effective screening test. Additionally, some patients may have contraindications to MRI testing (Doyle, 1999) and certain functional conditions, e.g. auditory neuropathy spectrum disorder (ANSD) may not be readily detected by imaging (Buchman *et al.*, 2006).

2.1.5 Conclusion

Accurate assessment of the auditory system is necessary for a number of important clinical applications. The ABR is an effective tool for evaluating the peripheral auditory nervous system and for estimating an individual's hearing thresholds. This is particularly useful for populations for whom traditional behavioural hearing tests are not possible (e.g. newborns). Accurate estimation of an individual's hearing threshold through ABR testing is reliant on the evoked potential being

accurately detected from noisy EEG recordings. ABR detection is the topic of the next chapter of this thesis where statistical ABR detection methods will be reviewed.

Chapter 3 ABR Detection

3.1 Recording the ABR

ABR detection is a very challenging task indeed. This is primarily the result of the signal of interest (the ABR) being much lower in amplitude than the background noise present in the measurement. This noise may arise from endogenous (physiological) sources, or exogenous (environmental) sources. Endogenous/physiological sources of noise include background EEG activity, myogenic (muscular) activity, cardiac activity, and ocular movements (largely myogenic plus movement of the corneo-retinal potential—a standing potential between the cornea and retina of the eye). External, exogenous, sources of noise in the recording include: mains interference, electrical equipment, and stimulus artefact (McLean, Scott and Parker, 1996). Stimulus artefact refers to the unwanted electrical activity picked up by the recording electrodes as a result of the electromagnetic activity of the transducer delivering acoustic stimuli to the ear (McLean, Scott and Parker, 1996). A large ABR may have an amplitude of 0.5 μV, whereas the background noise may be upwards of ±15 μV. The signal-to-noise ratio (SNR) of the ABR within the continuous EEG is therefore very low (around -35 to -23 dB depending on the stimulus level used—Chesnaye, 2019). Filtering to reduce energy in noisy frequency bands is useful, however, the benefits are limited by the significantly overlapping frequency spectra of the ABR signal and the background noise (Schimmel, Rapin and Cohen, 1974). The primary method of improving the SNR to allow ABR detection is through careful experimental technique to reduce noise and artefact at source, coupled with collecting multiple repeated measurements and applying coherent averaging.

3.1.1 Coherent Averaging

In order to improve the SNR of the ABR, multiple stimulus repetitions are performed. The EEG following each acoustic stimulus is recorded. These individual recordings are subsequently averaged together, decreasing the noise levels within the average whilst sparing the deterministic signal of interest (Dawson, 1954; Jewett, Romano and Williston, 1970). Each of these short EEG recordings is known as a ‘recording epoch’ and consists of a fixed period of EEG recorded following the onset of an auditory stimulus. The effectiveness of coherent averaging depends on a number of assumptions regarding the nature of evoked potentials (Elberling and Don, 1984):

1. The evoked potential is deterministic, i.e. the evoked potential signal is the same over each point in time for each recording epoch.

2. The noise is random, zero mean, independent and identically distributed (i.i.d.), and stationary.
3. The noise and the evoked potential signal are independent.

Whilst coherent averaging may still be performed if one or more of these assumptions are not met, the effectiveness of the technique will be reduced. In order to better visualise the structure of the ABR recordings it is helpful to consider the data as a matrix ($\mathbf{X}$) consisting of N rows of recording epochs (i.e. N acoustic stimuli) and M columns of sample points (i.e. each recording is of length M samples) (Chesnaye, 2019):

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ x_{N,1} & \cdots & \dots & x_{N,M} \end{bmatrix}$$

The individual recording epochs may be averaged together in the time-domain using the following equation provided by Lyons (2010):

$$\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i\,(t) \tag{3.1}$$

where $\bar{x}$ is the coherent average, and $x_i(t)$ is sample point (t) of the i^{th} recording epoch. As the temporal relation of the recording epochs to the onset of the auditory stimulus is constant, the averaging is said to be 'coherent'. With the evoked potential signal being deterministic, the noise will average out over a large number of recording epochs, leaving the averaged waveform with an improved SNR. A visual example of the effects of coherent averaging is presented below in Figure 3-1. Provided that the noise is stationary and that the number of recording epochs is large enough, the unweighted coherent average approximates the minimum mean square error (MMSE) estimate of the evoked potential signal (Schwartz and Shaw, 1975 in Hoke *et al.*, 1984). The noise is often not stationary and multiple alternative averaging methods have been proposed. Some of these methods will be explored in more detail in a subsequent chapter of this thesis (Chapter 5—Automated ABR Detection and Weighted Averaging).

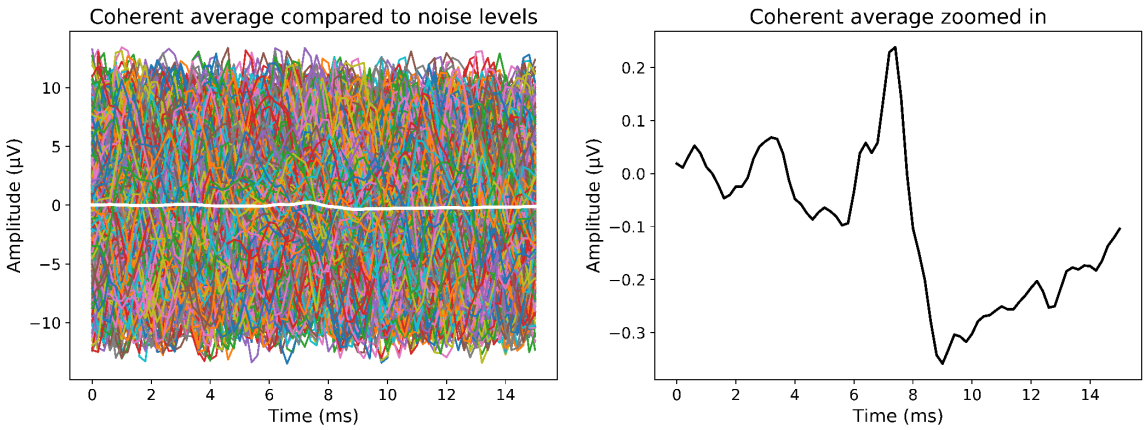

Figure 3-1 The averaged ABR waveform relative to background noise levels.
LEFT—The coloured traces each represent an individual recording epoch. The white line is the coherently averaged waveform, produced by averaging together all 3,000+ recording epochs. This shows clearly that the coherent average is very small in amplitude, compared to the signal from the individual epochs which comprise largely of noise.
RIGHT—This is the same waveform as the coherent average shown in the left plot (white), however, the y-axis display has been expanded to better view the ABR waveform morphology.

3.1.1.1 How Coherent Averaging Improves the SNR

The effect that averaging has on reducing noise whilst maintaining a constant evoked potential signal is relatively intuitive, however, it is important to understand the mathematical principles underlying this process. The structure of each epoch is demonstrated in an equation provided by Elberling & Don (1984):

$$x_i(t) = s_i(t) + v_i(t) \tag{3.2}$$

where *s* is the evoked potential of interest (the signal) and *v* is the background noise for sample point (t) within the i^{th} recording epoch, $x_i(t)$. The N recording epochs can be coherently averaged together across each sample point (Van Drongelen, 2018):

$$\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t) \tag{3.3}$$

where $\bar{x}$ is the coherent average of for sample point (t). Given that the evoked potential signal is considered to be deterministic and therefore identical for each sample point across recording epochs this can be rewritten as (Arar, 2019):

$$\bar{x}(t) \;=\; s(t) + \frac{1}{N}\sum_{i=1}^{N} v_i(t) \tag{3.4}$$

As the signal strength is theoretically unchanged with coherent averaging of a deterministic signal, the improvement of the SNR within the coherent average therefore arises as a result of a decrease in the denominator, i.e. the background noise level.

The SNR can be calculated as (Van Drongelen, 2018):

$$SNR = \frac{P_{signal}}{P_{noise}} \tag{3.5}$$

$$SNR_{dB} = 10\log_{10}\frac{P_{signal}}{P_{noise}} \tag{3.6}$$

with the average power (P) of a discrete time series being equal to its mean squared amplitude (Van Drongelen, 2018):

$$P = \frac{1}{K}\sum_{t=1}^{K} x(t)^2 \tag{3.7}$$

where K is the number of sample points in the time series. The variance of a signal is calculated similarly, however, with the mean value subtracted from each individual value:

$$\sigma^2 = \frac{1}{K}\sum_{t=1}^{K} (x(t) - \mu)^2 \tag{3.8}$$

For a zero-mean time-series, as EEG background noise is assumed to be, the power of the time-series will be equal to its variance (σ^2), which is the square of the standard deviation (σ). The variance of the noise is a measure of the intensity of the fluctuations in noise amplitude (Davenport Jr. and Root, 1987).

To understand how coherent averaging improves the SNR with increasing recording epochs (N), we must consider the effect of averaging on the variance of the noise (v) within the averaged waveform ($\bar{x}$). Van Drongelen (2018) and Arar (2019) provide the following equation:

$$\begin{aligned} Var\big(\bar{v}(t)\big) \;&=\; E[(\bar{v}(t) - \mu)^2] \\ =\; E[\bar{v}(t)^2 - 2\bar{v}(t)\mu + \mu^2\,] \;&=\; E[\bar{v}(t)^2] - 2\mu E[\bar{v}(t)] + \mu^2 \;=\; E[\bar{v}(t)^2] - \mu^{\,2} \\ =\; E\left(\left[\frac{1}{N}\sum_{i=1}^{N} v_i(t)\right]^2\right) - \mu^2 \;&=\; E\left[\frac{1}{N^2}\sum_{i=1}^{N} v_i(t)\sum_{j=1}^{N} v_j(t)\right] - \mu^2 \end{aligned} \tag{3.9}$$

where E is the expectation operator and μ is the mean value of the noise within the coherently averaged background noise, $\bar{v}$, for sample point (t). Note that μ is the true mean value of $\bar{v}(t)$, and so $E[\mu] = \mu$ (Van Drongelen, 2018). Additionally, $E[\bar{v}(t)] = \mu$, allowing the simplifications which occur in the second line of Equation 3.9 (Van Drongelen, 2018). The two summations in the final expression of Equation 3.9 represent N^2 possible combinations of i and j (Van Drongelen, 2018). To better understand what's happening, we may separate this expression into the $N(N-1)$ terms in the expression where $i \neq j$, and the N terms where $i = j$, in an equation provided by Van Drongelen (2018):

$$Var(\bar{v}(t)) = \frac{1}{N^2} \underbrace{\overbrace{\sum_{i=1}^{N} E[v_i(t)^2]}^{N\ terms}}_{for\ i=j} + \frac{1}{N^2} \underbrace{\overbrace{\sum_{i=1}^{N}\sum_{j=1}^{N} E\big[v_j(t)v_i(t)\big]}^{N(N-1) terms}}_{for\ i\neq j} - \mu^2 \tag{3.10}$$

On the basis that the mean value of the noise (μ) is assumed to be equal to zero, μ^2 may be disregarded. A further assumption made regarding the properties of the noise is that the noise is independent between recording epochs (i.e. the 'independence assumption') (Van Drongelen, 2018). For the $N(N-1)$ terms in the expression where $i \neq j$, two different i.i.d. random variables with an expected value of zero are multiplied together producing a value of zero (Arar, 2019):

$$E\big[v_i(t)v_j(t)\big] = E[v_i(t)]E\big[v_j(t)\big] = 0 \quad for\ i \neq j \tag{3.11}$$

The second component of the final expression in Equation 3.10 is therefore assumed to be zero and can be removed. For the remaining N terms in the expression where $i = j$, Arar (2019) provides the following equation:

$$E\big[v_i(t)v_j(t)\big] = E[v_i(t)^2] = \sigma_v(t)^2 \quad for\ i = j \tag{3.12}$$

where ${\sigma_v}^2$ is the variance of the background noise, v. Combining Equations 3.10, $3.11$and 3.12, for the N terms where $i = j$, we can observe the effect of averaging on the noise within the coherent average in the equation provided by Arar (2019):

$$Var(\bar{v}(t)) = \frac{1}{N^2} \underbrace{\overbrace{\sum_{j=1}^{N} E\big[v_j(t)^2\big]}^{N\ terms}}_{for\ i=j} = \frac{1}{N^2} N\sigma_v(t)^2 = \frac{\sigma_v(t)^2}{N} \tag{3.13}$$

The variance of the noise therefore decreases by a factor of *N* with coherent averaging.

Now that the effect of averaging on the noise level has been considered, we move on to observing the effects of averaging on the SNR. Based on Equations 3.5 and 3.13, we can observe that the variance of the noise in the averaged waveform decreases by a factor of N with coherent averaging, whilst the SNR increases by the same factor for EEG where a response is present:

$$SNR_{\bar{\mathbf{x}}} = N \frac{P_{\mathbf{s}}}{{\sigma_{\mathbf{v}}}^2} \tag{3.14}$$

Some authors describe the SNR in terms of the amplitude of the signal in relation to the amplitude of the noise (Hall, 2007; Lyons, 2010). For clarity, this version of calculating the SNR will be referred to as the Signal-amplitude-to-noise ratio (SANR). As the maximum amplitude of the noise is a much less stable estimate of the noise level compared to the standard deviation and both are linearly related, the SANR may be calculated as (Lyons, 2010):

$$SANR_{\bar{\mathbf{x}}} = \frac{A_{\mathbf{s}}}{\sigma_{\bar{\mathbf{v}}}} \tag{3.15}$$

where A_S represents the amplitude of the coherently averaged ABR (from the peak of wave V to the trough of SN10) and $\sigma_{\bar{\mathbf{v}}}$ represents the standard deviation of the noise, within the coherent average. Whereas the variance of the noise decreases by a factor of *N* with coherent averaging (Equation 3.13), the denominator of the SANR is the standard deviation of the noise (the square root of the variance). The SANR therefore improves by a factor of $\sqrt{N}$ for EEG where a response is present:

$$SANR_{\bar{\mathbf{x}}} = \sqrt{N} \frac{A_{\mathbf{s}}}{\sigma_{\mathbf{v}}} \tag{3.16}$$

where $\sigma_{\mathbf{v}}$ is the standard deviation of the noise within a single recording epoch/the continuous EEG. The above theory regarding how coherent averaging improves the SNR and SANR is demonstrated experimentally, through simulation, in Figure 3-2.

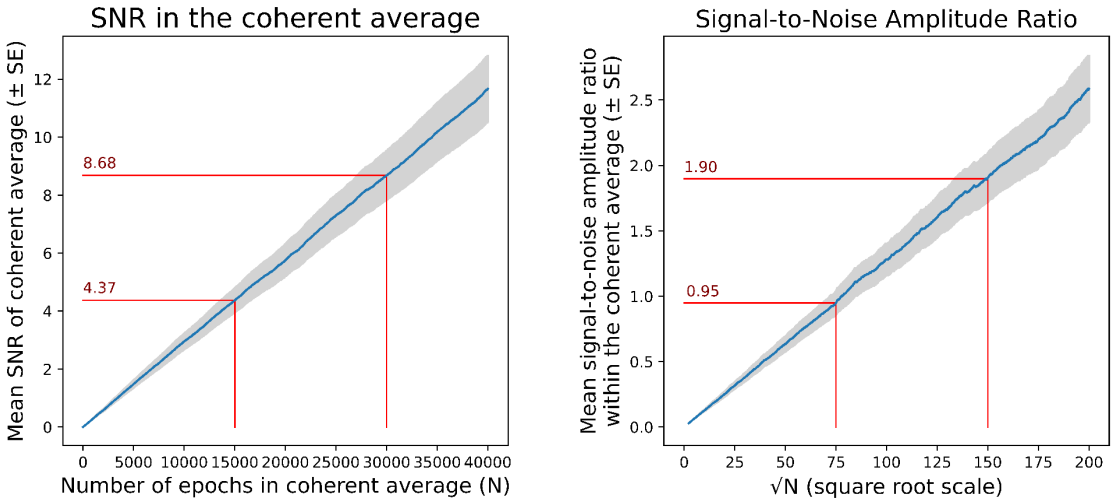

Figure 3-2 SNR (Equation 3.5) improves with the number of recording epochs in the coherent average. Both the left and right graphs are based on the same simulated ABR data; a fixed ABR template has been added to 40,000 epochs of noise drawn randomly from a Gaussian distribution centred at zero, repeated 100 times. The LEFT graph shows how SNR as a ratio of signal power to noise power increases by a factor of N epochs with coherent averaging. Note the approximately linear relationship between the SNR of the coherent average and the number of recording epochs, as predicted by Equation 3.14. The RIGHT graph demonstrates how the signal-to-noise amplitude ratio increases by a factor of $\sqrt{N}$ epochs with coherent averaging, as predicted from Equation 3.16—note the square root scale on the x-axis.

3.2 ABR Detection Methods

3.2.1 Signal Detection

The goal of ABR detection methods is to successfully differentiate between EEG recordings containing a signal (the ABR) and EEG recordings containing just noise (no response). As a result of the high noise levels present in EEG recordings and the low amplitude of the ABR evoked potential signal, differentiating recordings containing a response from recordings containing solely noise is extremely challenging. As there are typically considered to be two possible detection outcomes (‘response present’ or ‘response absent’), this task is considered to be a binary classification problem (Figure 3-3). With clinical applications in mind, it is helpful to be aware of uncertainty and so some detection procedures include the possibility of a third ‘inconclusive’ class (British Society of Audiology, 2019c; McKearney and MacKinnon, 2019).

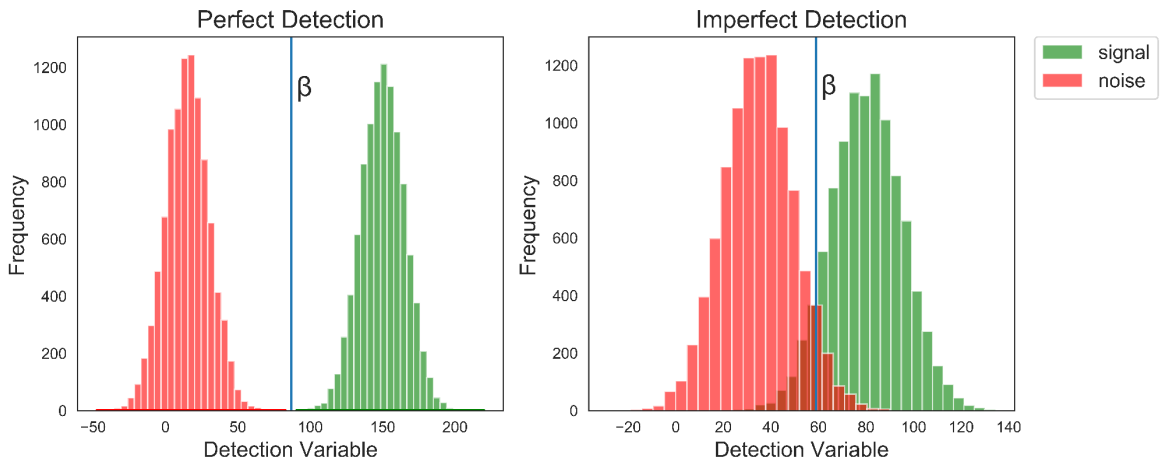

Figure 3-3 Signal detection—binary classification. The LEFT graph shows the ideal scenario whereby the detection method is able to differentiate fully between the EEG containing an evoked potential signal and EEG containing no signal when using an appropriately chosen decision criterion (β). In this example, there are no false alarms or misses, as the detection criterion perfectly separates the signal and the noise. The RIGHT graph shows the more commonplace scenario whereby the detection variable is not entirely able to differentiate between the case where the recording contains an evoked response or only the noise. A detection criterion will have to be chosen to optimise the detection performance for the specific application. Cases where a ABR signal is present, but the detection variable is below the detection criterion (β) represent misses as the response is not detected (Anderson, 2015). Cases where there is no response present, but the detection variable is above the detect ion criterion (β) represent false alarms (Anderson, 2015). This figure was redrawn, with changes made (not all graphs were included, and the data used in the graphs was based on simulated data produced by the present author), from a figure by Anderson, N. D. (2015) 'Teaching signal detection theory with pseudoscience', Frontiers in Psychology. Frontiers Research Foundation, 6(JUN), p. 762. Available at: https://doi.org/10.3389/fpsyg.2015.00762, under the terms of the CC BY 4.0 licence under which the work was published.

3.2.2 Visual Inspection

ABR detection in clinical practice is largely based on visual inspection (Vidler and Parker, 2004; British Society of Audiology, 2019c). Clinicians will typically examine pairs or single coherently averaged waveforms across a range of stimulus levels (Figure 3-4). Information from these waveforms such as the reproducibility, latency and amplitude of any response, and morphology will be used to estimate the individual's ABR threshold (British Society of Audiology, 2019c).

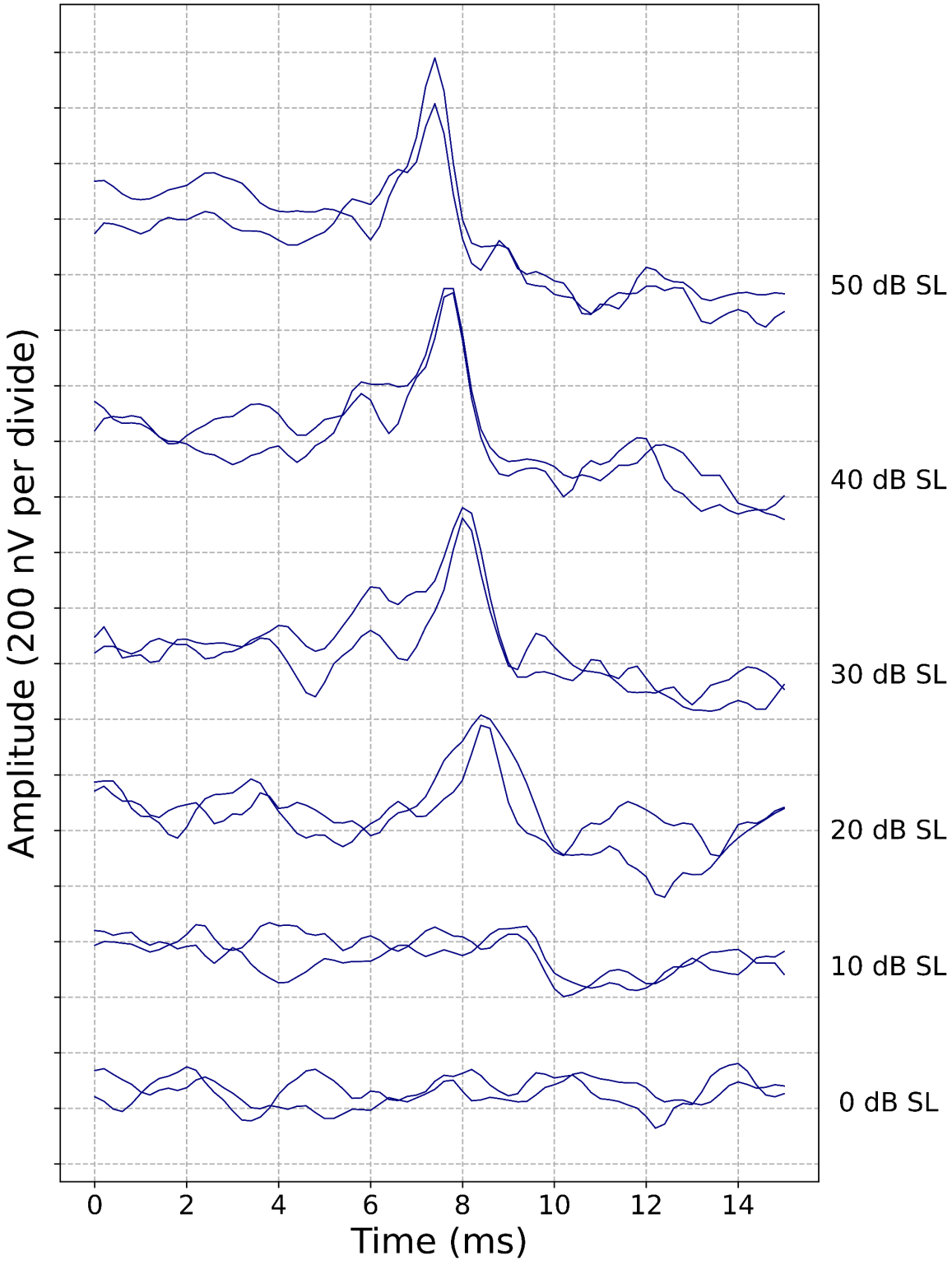

Figure 3-4 Example of ABR data used for threshold detection. Repeat recordings have been performed for stimulus levels from 0 to 50 dB SL in steps of 10 dB. Using the using the British Society of Audiology (2019c) guidelines to aid interpretation, clear responses appear to be present for stimulus levels down to and including 10 dB SL. At 10 dB SL the response size is small (~220 nV), and background noise is present. However, the response amplitude is greater than three times the background noise

amplitude, as estimated by the average absolute difference between the two waveforms (57 nV). At 0 dB SL, the waveforms are not appropriately flat, nor is the background noise level below 25 nV. The waveforms at 0 dB SL are therefore 'inconclusive'. The threshold in this case is considered to be <= 10 dB SL.

Visual interpretation requires skill and training and interpretation is known to vary significantly, even amongst experienced clinicians (Cohen *et al.*, 1971; Vidler and Parker, 2004). Cohen *et al.* (1971) remark that 'while the human visual system is an extremely sensitive one for pattern recognition, it is also highly subjective and, accordingly, visual scoring of the [auditory evoked response] leads to inconsistent results'. Vidler and Parker (2004) designed a study whereby 16 experienced clinical audiologists (with an average of 8-years' experience in ABR interpretation) were asked to take part in a simulation of ABR recording and interpret the results. Pre-recorded ABR data were made available in a simulation which the clinicians controlled, with clinicians being able to choose the stimulus levels used and the number of recording epochs obtained. The clinicians were tasked with running the ABR recording simulation in order to make a decision regarding the participant's ABR threshold across 12 ABR datasets. Interpretation varied significantly; for nine of the 12 ABR datasets the difference between the lowest and highest estimated ABR threshold was ≥40 dB nHL. This level of variability could significantly affect the diagnosis of an individual's hearing status and in turn their management, including any amount of amplification prescribed. There are some mitigating factors that may account in part for the variability in performance observed: the simulation being constrictive compared to real-life ABR recording, and the simulation software being unfamiliar to the clinicians (Vidler and Parker, 2004). These factors are unlikely to account wholly for the variability observed in expert threshold estimation and so the results bring concern regarding the ability of human experts to reliably interpret the ABR through visual inspection.

3.2.3 Statistical Detection Methods

Some clinical AEP software provides clinicians with statistical response confidence measures. These confidence measures provide a quantitative estimate of the magnitude of the response and/or a significance level (p-value) for a given hypothesis test, to help clinicians determine if a response is present. These include the Fsp (Elberling and Don, 1984) and Fmp (Martin *et al.*, 1994) which both compare the variance of the coherent average (the estimated signal) to the variance of the estimated noise level (British Society of Audiology, 2019c). The British Society of Audiology guidelines (2019c) advise that these confidence measures may be used to determine whether a recording satisfies the criterion of having a SNR of ≥ 3:1 in order to be deemed a 'clear response'. However, the other 'clear response' criteria relating to the ABR morphology and waveform

replicability rely upon visual inspection. Additionally, the Fsp/Fmp may not be used to determine if a response is absent. Interpretation therefore remains reliant upon visual inspection. Some AEP software also include a measure of the residual noise level within the coherent average. The British Society of Audiology guidelines (2019c) recommend that residual noise measures may guide clinicians' decisions as to when to stop a recording. However, in order for EEG to be considered 'response absent', the EEG must still fulfil all of the 'response absent' criteria which rely wholly upon visual inspection (British Society of Audiology, 2019c). In conclusion, whilst there are tools available to help clinicians in their interpretation of EEG recordings, current national guidelines are clear that interpretation should be based on visual interpretation by trained clinicians. Improvements in the performance of automated ABR detection methods may have the potential to shift emphasis away from visual inspection if they are deemed more reliable. The next sections of this chapter will provide an overview of the more prominent statistical ABR detection methods in the literature. These have been previously described and compared by Chesnaye *et al.* (2018;2019) for readers who seek additional detail. The work by Chesnaye *et al.* (2018;2019) served as inspiration for inclusion of the following statistical detection methods in this review and in the study presented in Chapter 4.

3.2.3.1 The Fsp and the Fmp

The Fsp and Fmp are very closely related and so will be considered together in this section. Originally described by Elberling and Don in 1984, the Fsp provides a measure of the likelihood that the null hypothesis ('response absent') can be rejected. The Fsp is a calculation of the ratio of the estimated variance of the estimated evoked potential signal (the coherent average) over the variance of the estimated background noise levels within the coherent average. The equation for calculating the Fsp is provided by Elberling and Don (1984) as follows:

$$Fsp = \frac{Var(\bar{\mathbf{x}})}{Var(\overline{\mathbf{sp}})} \tag{3.17}$$

with $\bar{\mathbf{x}}$ being the coherent average and $\overline{\mathbf{sp}}$ being calculated as:

$$Var(\overline{\mathbf{sp}}) = \frac{Var(\mathbf{X}_{.j})}{N} \tag{3.18}$$

where $\mathbf{X}_{.j}$ is a column vector of sample points down a single *chosen* column (j) of the ensemble matrix $\mathbf{X}$. Equations 3.17 and 3.18 may therefore be combined to be written as:

$$Fsp = \frac{Var(\bar{\mathbf{x}})}{\frac{1}{N}Var(\mathbf{X}_{.j})} \tag{3.19}$$

The Fsp is a variance-ratio test (*F*-test). The denominator is $\frac{1}{N}$ the estimated variance of the continuous EEG, as the variance of the noise within the coherent average is expected to reduce by a factor of N with coherent averaging, as discussed in section 3.1.1.1, under the assumption of Gaussianity. The numerator of the Fsp equation provides an estimate of the variance of the ABR signal intertwined with the variance of the averaged background noise (Elberling and Don, 1984):

$$Var(\bar{\mathbf{x}}) = Var(\mathbf{s}) + Var(\bar{\mathbf{v}}) + 2 \cdot Cov(\mathbf{s}, \bar{\mathbf{v}}) \tag{3.20}$$

If no ABR signal is present, then the Fsp would simply be a ratio of the estimated variance of the averaged background noise over the estimated variance of the averaged background noise and be expected have a value of ~1. In the case of an evoked potential signal being present, the value of the numerator is expected to be greater than that of the denominator, leading to an expected Fsp value of >1. As a variance-ratio test, the Fsp is expected to follow an *F*-distribution with the degrees of freedom v_1 relating to the independence between samples in the coherent average (numerator, see below) and $v_2 = N - 1$ degrees of freedom (Elberling and Don, 1984), assuming independence between the background EEG noise between recording epochs (Chesnaye, 2019). If the average EEG background noise were i.i.d. random variables, then v_1 would be equal to $M - 1$, with M being the number of samples in the coherent average (Elberling and Don, 1984). However, EEG noise does not meet this assumption, with narrow bands of dominant frequencies leading to reduced independence between samples within recording epochs and therefore a reduced number of degrees of freedom (v_1) (Elberling and Don, 1984; Chesnaye *et al.*, 2018). The degrees of freedom of the numerator in the EEG noise is unknown and difficult to estimate. Based on empirical data, Elberling and Don (1984) recommended a conservative value for the degrees of freedom of $v_1 = 5$, as most EEG noise will have at least this number of degrees of freedom for v_1. Choosing a conservative value for v_1, prevents the false positive rate from being too high (i.e. a conservative detector of evoked responses) as most EEG noise will have a greater number of degrees of freedom than this. The value v_2 relates to the degrees of freedom for the denominator. As the column vector $\mathbf{sp}$ comprises N sample points which are assumed to be independent, i.e. independence between recording epochs, v_2 is said to be equal to the number of recording epochs in the ensemble (N).

The relationship of the Fsp statistic to the SNR of the averaged recording is provided by Elberling and Don (1984):

$$Fsp = \left(SNR + 1 + 2 \cdot R(\mathbf{s}, \bar{\mathbf{v}}) \cdot \sqrt{SNR}\right) \cdot \frac{Var(\bar{\mathbf{v}})}{Var(\overline{\mathbf{sp}})} \tag{3.21}$$

The Fsp is influenced by previous work. Schimmel, Rapin and Cohen (1974) proposed a variety of methods for objectively evaluating evoked potential data, including a power ratio of the mean

post-stimulus interval to the mean pre-stimulus interval which 'resembles the variance statistic F'. Wong and Bickford (1980) adapted this measure to be the ratio of the variance of the coherent average to the variance of the background noise, estimated using the ± reference technique described by Schimmel (1967). The ± reference is a form of average calculated by alternate addition and subtraction of alternate recording epochs, cancelling out the response if present (Schimmel, 1967).

The Fmp, proposed by Martin *et al.*, 1994, involves an alteration to the denominator of Fsp equation (Equation 3.17). Rather than estimating the level of the background noise by calculating the variance down a single column across recording epochs, the variance is calculated down multiple columns of points before being averaged together. Whilst first described verbally by Martin *et al.* (1994), the equation for calculating the Fmp is provided by Cebulla, Stürzebecher and Wernecke (2000):

$$Fmp = \frac{Var(\bar{\mathbf{x}})}{\frac{1}{N}\left(\frac{1}{Q}\sum_{i=1}^{Q} Var\left(\mathbf{X}_{\cdot,i}\right)\right)} \tag{3.22}$$

Where Q is the number of chosen single point columns that are used to estimate the noise level and $\mathbf{X}_{\cdot,i}$ is the i^{th} chosen column of the ensemble matrix $\mathbf{X}$. The number of single point columns used for analysis (Q) may be multiple or all columns available, i.e. $1 < Q \leq M$. If the number of single point columns used (Q), is equal to 1, the *F* statistic calculated would be equal to the Fsp as provided by Equation 3.17, assuming the same single point column were used.

3.2.3.2 Hotelling's T^2 Test

Another statistical detection method which has been effectively applied to AEP detection is the one-sample Hotelling's T^2 test (HT2) (Picton *et al.*, 1987; Valdes-Sosa *et al.*, 2009). The HT2 test was originally described by Hotelling in 1931, and is a multivariate extension of the one-sample Student's *t*-test (Student, 1908). The one-sample Student's *t*-test explores the hypothesis that a population mean is significantly different to a known or theoretical mean, whereas the one-sample HT2 test explores the hypothesis that a multivariate sample mean (containing two or more variables) is significantly different to an expected (known or theoretical) multivariate mean (King and Eckersley, 2019). In the case of ABR detection in the time domain we explore the hypothesis that the means within the coherent average are significantly different from a hypothesised mean value of zero (Elberling and Don, 1984). The Hotelling's T^2 test may also be applied to frequency domain ABR data (Chesnaye *et al.*, 2018). The Hotelling's T^2 is calculated as follows (Hotelling, 1931; King and Eckersley, 2019):

$$T^2 = N\,(\bar{\mathbf{x}} - \boldsymbol{\mu})^T\,\mathbf{C}^{-1}\,(\bar{\mathbf{x}} - \boldsymbol{\mu}) \tag{3.23}$$

where $\bar{\mathbf{x}}$ is the vector of sample means, $\boldsymbol{\mu}$ is the vector of expected mean values (under the null hypothesis, i.e. zero in the case of evoked potentials), $\mathbf{C}$ is the sample covariance matrix, and N is the sample size of the population (equal to the number of recording epochs).

The test statistic for the Hotelling's T^2 test may be calculated using the T^2 value produced by Equation 3.23, using the following equation (King and Eckersley, 2019):

$$F = \frac{N - p}{p(N - 1)} T^2 \tag{3.24}$$

where p is the number of different variables, which will be elaborated upon in the next section of this chapter. This *F* statistic is known to follow an *F*-distribution with $v_1 = p$ and $v_2 = (N - p)$ degrees of freedom (King and Eckersley, 2019). The critical value for rejecting the null hypothesis may therefore be obtained from this *F*-distribution.

3.2.3.2.1 The Number of Voltage Means

An important parameter value to be chosen when applying the HT2 test to ABR data in the time domain is the number of different variables, i.e. the value of p from Equation 3.24 (Golding *et al.*, 2009; Van Dun, Dillon and Seeto, 2015; Chesnaye *et al.*, 2018; Chesnaye, 2019). The ABR data consist of an ensemble matrix of N rows of recording epochs by M columns of sample points. The HT2 test may be applied to all of the M columns, i.e. such that the number of variables $p = M$. However, it is known that including too many variables reduces the sensitivity of the test as the likelihood of chance affecting the outcome increases (Golding *et al.*, 2009). In the time domain, the M sample points in each recording epoch may be compressed into a smaller number of data-bins by averaging groups of adjacent sample points, so that $1 < p < M$ (Golding *et al.*, 2009). These will be referred to as voltage means (Chesnaye *et al.*, 2018). The number of voltage means should neither be so high that each additional bin adds little/no additional information, nor be so low that information is lost as a result of over-compression (Golding *et al.*, 2009; Van Dun, Dillon and Seeto, 2015; Chesnaye *et al.*, 2018). In relation to the ABR, using too few voltage means would lead to peaks and troughs of the ABR morphology falling within the same data-bin, cancelling each other out and resulting in the corresponding voltage means being close to zero (Van Dun, Dillon and Seeto, 2015; Chesnaye, 2019). A previous study by Chesnaye *et al.*, 2018b, found the optimal number of voltage means to be 25 for the ABR, with similar levels of performance using any number of voltage means between 20 and 40. The optimal number of voltage means will be data dependant relying on factors such as the length of the analysis window and the sampling rate.

3.2.3.3 The q-sample Uniform Scores Test

The q-sample uniform scores test was described by Mardia in 1972 and was first applied to AEP detection by Stürzebecher, Cebulla and Wernecke in 1999. The q-sample uniform scores test evaluates the uniformity of q samples of phase distributions (Stürzebecher, Cebulla and Wernecke, 1999) (Figure 3-5). First a fast Fourier transform (FFT) is applied to each of the N recording epochs in the ensemble. The phase angle of each of q Fourier components in each recording epoch are then calculated, concatenated into a single sequence (of length $N \times q$), and then ranked. The equation for calculating the test statistic (W) is provided by Stürzebecher, Cebulla and Wernecke (1999):

$$W = \frac{2}{N} \cdot \sum_{k=1}^{q} \left({C_k}^2 + {S_k}^2 \right) \tag{3.25}$$

with:

$$C_k = \sum_{i=1}^{N} \cos \beta_{ik} \; ; \qquad S_k = \sum_{i=1}^{N} \sin \beta_{ik} \tag{3.26}$$

and with β_{ik} being the uniform scores of the phase ranks:

$$\beta_{ik} = \frac{2 \cdot \pi \cdot r_{ik}}{N \cdot q} \tag{3.27}$$

where r_{ik} is the phase rank of the k^{th} sample within the i^{th} recording epoch. The test statistic W is expected to follow a χ^2 distribution with $2(q-1)$ degrees of freedom (Stürzebecher, Cebulla and Wernecke, 1999) under the null hypothesis of a uniform phase distribution.

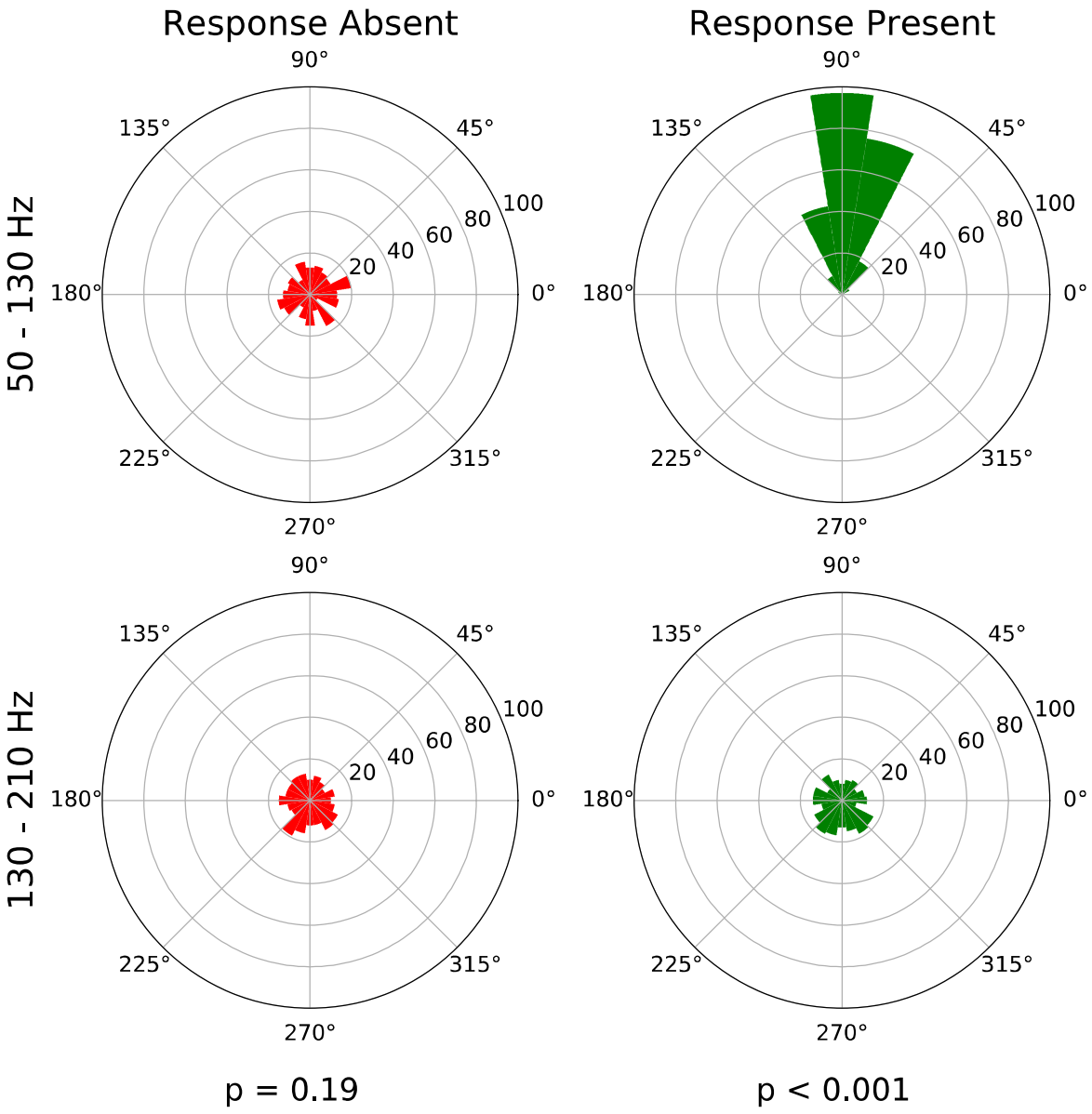

Figure 3-5 Analysing the uniformity of phase distributions. This figure shows circular histograms of phase angles, for simulated white noise ('Response Absent'), and white noise containing a 100 Hz sine wave response signal ('Response Present'). Two frequency bins (50–130 and 130–210 Hz) are evaluated ($q = 2$). The q-sample uniform scores test evaluates the uniformity of the phase angle distributions in combination across the q samples. The 'Response Present' histogram for the 50–130Hz frequency bin shows a highly non-uniform distribution, resulting in a low q-sample uniform scores test p value. The distributions of the phases in the two frequency bins of the 'Response Absent' data are uniformly distributed, resulting in a large p value.

3.2.3.4 Modified Versions of the Test

The original q-samples uniform scores test considers only the phase angles, neglecting the amplitude of the frequency components which potentially contains information which may aid

AEP detection (Stürzebecher, Cebulla and Wernecke, 1999). Various modified versions of the test exist which include information regarding the phase angles, spectrum amplitudes, or both (Cebulla, Stürzebecher and Elberling, 2006). The original version uses the phase angle ranks only (Mardia, 1972). Version 1 uses the raw phase angle values only (Cebulla, Stürzebecher and Elberling, 2006). Version 2 uses the ranks of both the phase angles and the spectral amplitudes (Stürzebecher *et al.*, 1996; Stürzebecher, Cebulla and Wernecke, 1999). Version 3 uses the raw values of the phase angles and the ranks of the spectral amplitudes (Cebulla, Stürzebecher and Elberling, 2006). Version 4 uses the raw values of both the phase angles and the spectral amplitudes (Cebulla, Stürzebecher and Elberling, 2006). This version numbering is consistent with that used by Cebulla, Stürzebecher and Elberling (2006).

There are a number of parameters which may be used to optimise these detection methods, including (Chesnaye *et al.*, 2018; Chesnaye, 2019):

- The length and position of the analysis window.
- Which spectral components to include.
- Increasing the number of frequency bins using zero padding.

In a comparison study by Stürzebecher, Cebulla and Wernecke (1999), the version 2 modified q-sample uniform scores test was found to be more effective than the original q-sample test and the q-sample analogue of the Watsons U^2 test (Maag, 1966). In a study by Chesnaye *et al.* (2018; 2019), version 2 of the q-sample uniform scores test was found to be more effective at detecting the ABR than version 4, using simulated data (Figure 3-6).

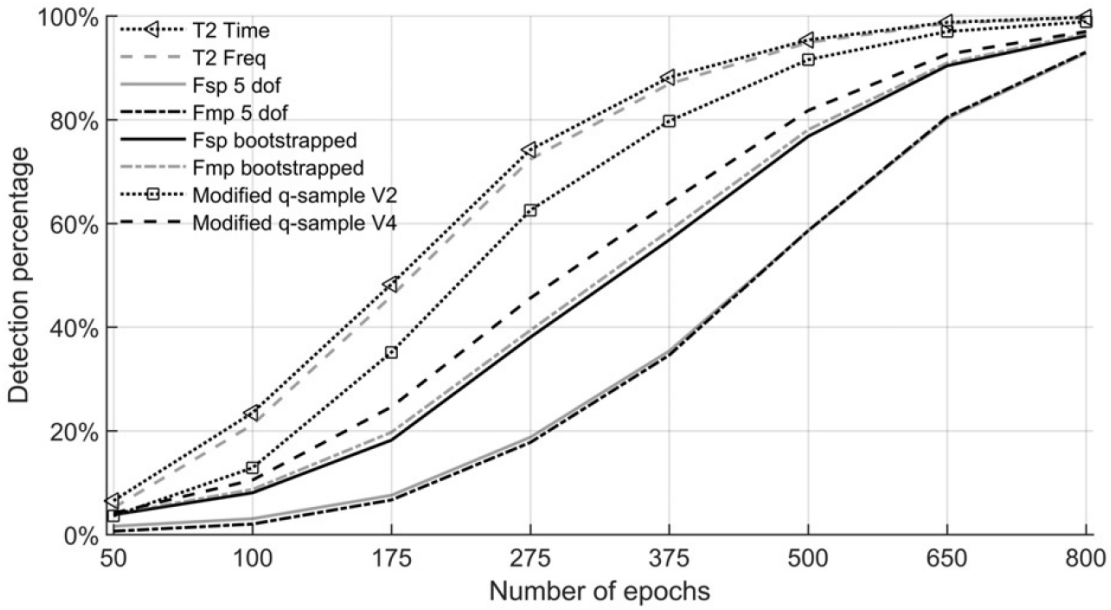

Figure 3-6 Comparison of ABR detection methods. Figure from Chesnaye *et al.*, 2018, reproduced with permission from Taylor and Francis (www.tandfonline.com).

3.3 The Bootstrap Technique

Many of the aforementioned ABR detection methods produce statistics which are expected to follow distributions that can be derived from mathematical theory under certain assumptions: the Fsp/Fmp (when assuming the degrees of freedom for the numerator), Hotelling's T^2 test, and the original q-sample uniform scores test. This allows the significance of the test statistic to be evaluated and a *p* value to be obtained. For other detection methods (e.g. the modified versions of the q-sample uniform scores tests), the mathematical derivations are difficult or intractable. However, critical values for rejecting the null hypothesis ('response absent') may be estimated by simulation (Feiveson, 2002). In other cases, the sampling distribution of the test statistic is strongly dependant on the characteristics of the individual recording and therefore cannot be estimated *a priori*. This would also be the case for new detection methods developed based on machine learning techniques. An effective method of obtaining a *p* value for these methods is to use the bootstrap (Efron and Gong, 1983; Lv, Simpson and Bell, 2007; Chesnaye *et al.*, 2018; Chesnaye, 2019).

For each ensemble of ABR data being evaluated, the first step is to calculate the test statistic in the conventional manner using the coherently arranged ensemble (a matrix of N recording epochs by M EEG sample points) (Lv, Simpson and Bell, 2007). This is demonstrated in Figure 3-7, where the Fsp value was calculated to be 2.15.

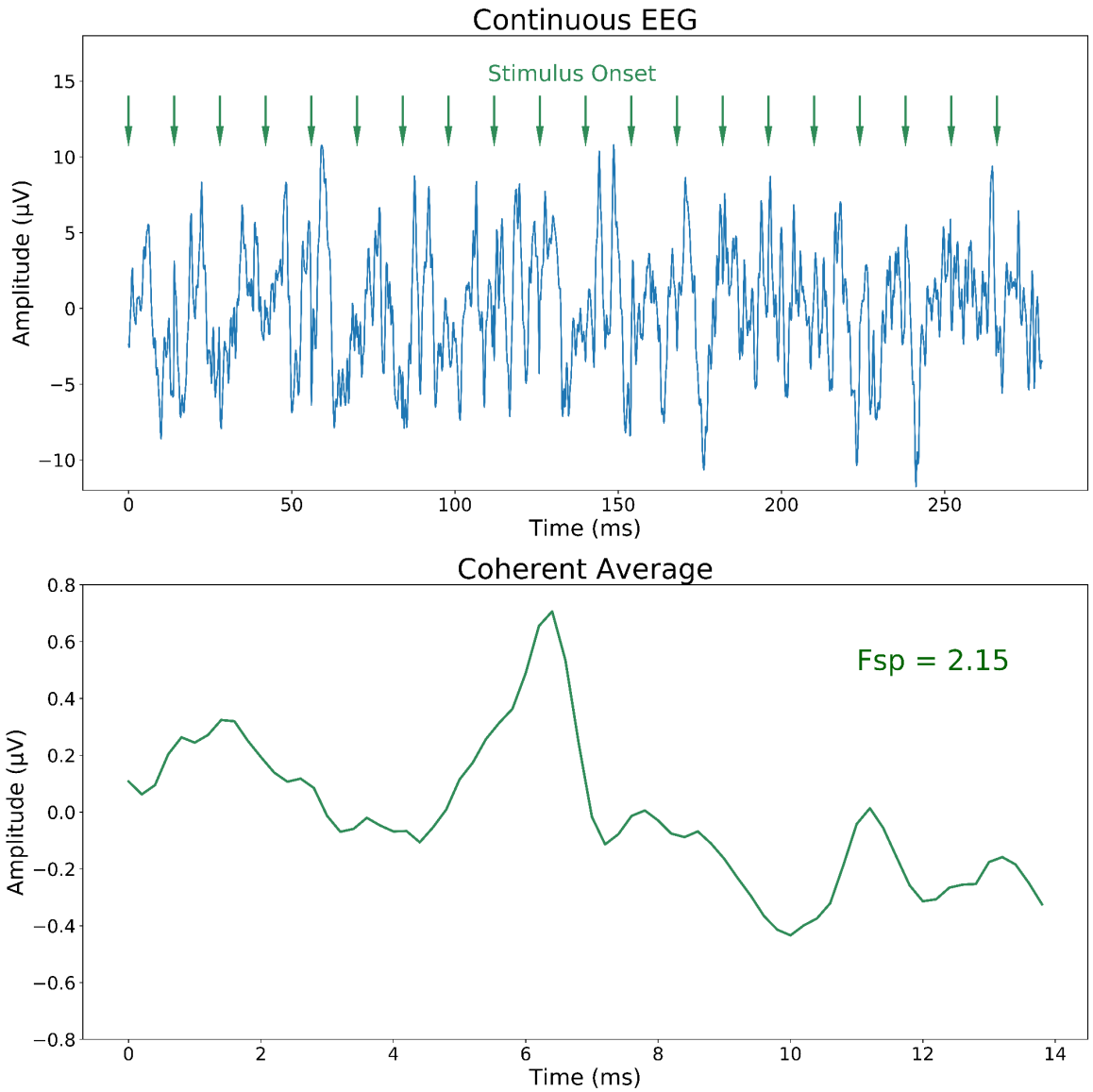

Figure 3-7 Calculation of the test statistic from the coherently arranged data. Each recording epoch, starting at each stimulus onset (green arrows), was aligned into a coherently arranged ensemble. The Fsp value was then calculated based on the coherently arranged data. Note, for the sake of clarity, the number of recording epochs (N) in the top figure is limited to 20, however, the coherent average and test statistic were calculated from all of the recording epochs in the ensemble ($N = 450$ in this case).

The next step is to determine the significance of the original test statistic by estimating the null distribution (i.e. the sampling distribution under the null hypothesis) of the data using the bootstrap. For each bootstrap sample, N random locations within the continuous EEG data are selected, irrespective to the stimulus onset timings (Lv, Simpson and Bell, 2007). Continuous sections of M discrete-time EEG samples are taken from each of the N randomly selected start points and used to construct an 'incoherent' ensemble. The ensemble is said to be incoherent as the starting points of each section of continuous EEG data are aligned irrespective of the stimulus

onset (Lv, Simpson and Bell, 2007). When these incoherently aligned continuous sections of EEG data are averaged together, the evoked potential signal is disrupted and cancelled out, thereby reflecting the properties of the null data ('no response') Figure 3-8. As an extra precaution, alternate EEG sections in the bootstrap sample ensemble may be inverted to help ensure that the evoked potential is fully disrupted in order to estimate the null distribution without bias (Chesnaye, 2019).

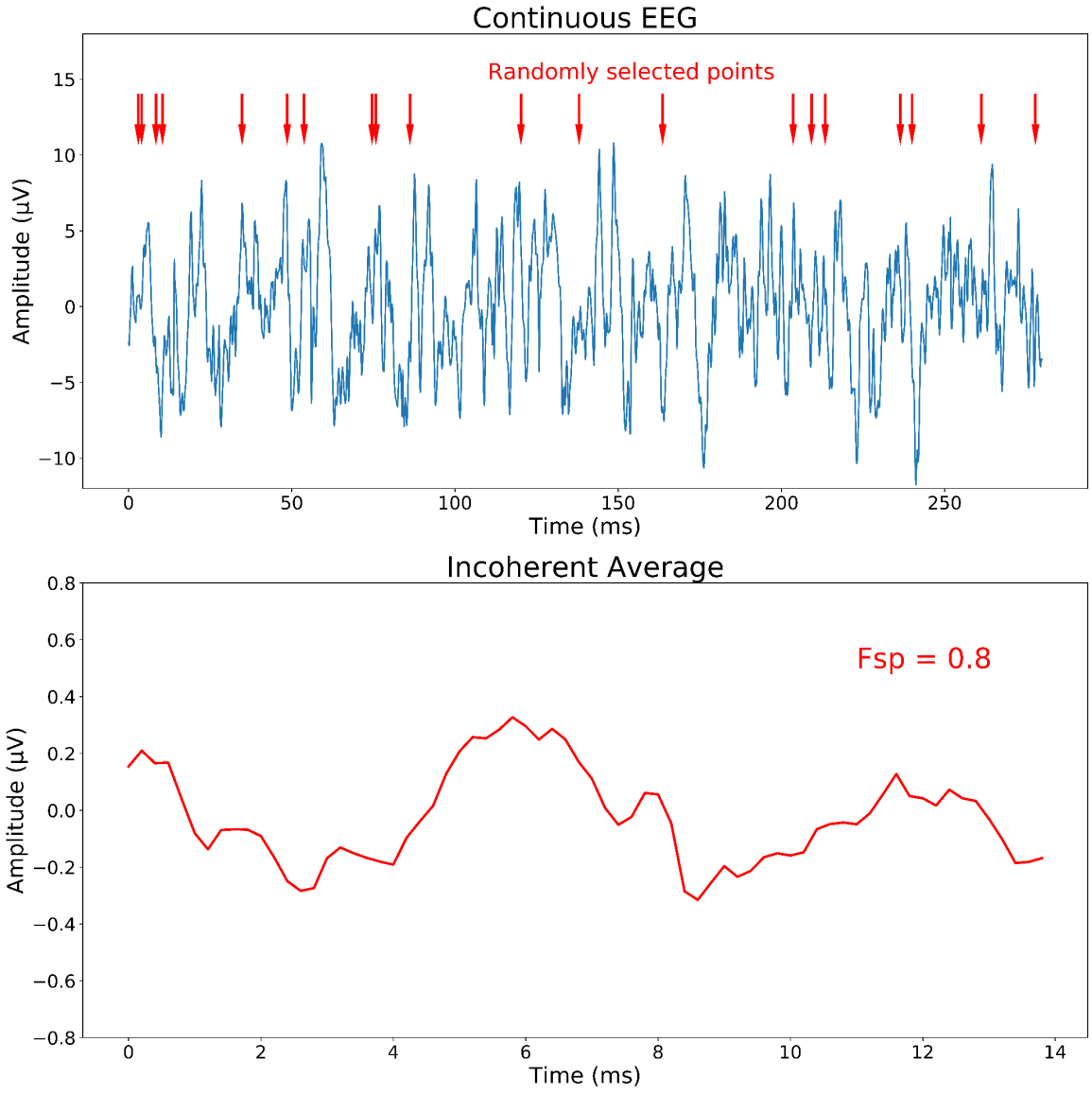

Figure 3-8 A bootstrap sample generated by selecting N sections of EEG from random locations within the continuous EEG.

This process is repeated for each of the bootstrap samples, with random start points chosen for the continuous EEG sections in each bootstrap sample, generating an estimated null distribution (Figure 3-9). The significance level of the original test statistic may then be calculated by evaluating its position within the bootstrapped null distribution (Lv, Simpson and Bell, 2007; Chesnaye *et al.*, 2018; Chesnaye, 2019). It should be noted that the bootstrap approach tests if

the test statistic calculated from the coherently arranged data differs significantly to the test statistic calculated from the data when recording epochs are not synchronised with the stimulus (Chesnaye, 2019). This approximates, but is not identical to, the null hypothesis of no response being present as the responses are still present (albeit asynchronously) when calculating the bootstrapped sampling distribution using incoherent averaging (Chesnaye, 2019). For evoked potentials such as the ABR, where the response amplitude is small, this is not expected to have a meaningful impact, with the bootstrapped null distribution closely approximating the true null distribution (Chesnaye, 2019). Further measures such as subtracting the coherent average from the recording epochs before performing the bootstrap, or inverting alternate EEG sections forming the bootstrapped ensemble may serve to allow the bootstrapped condition to better mimic the target null hypothesis (Chesnaye, 2019).

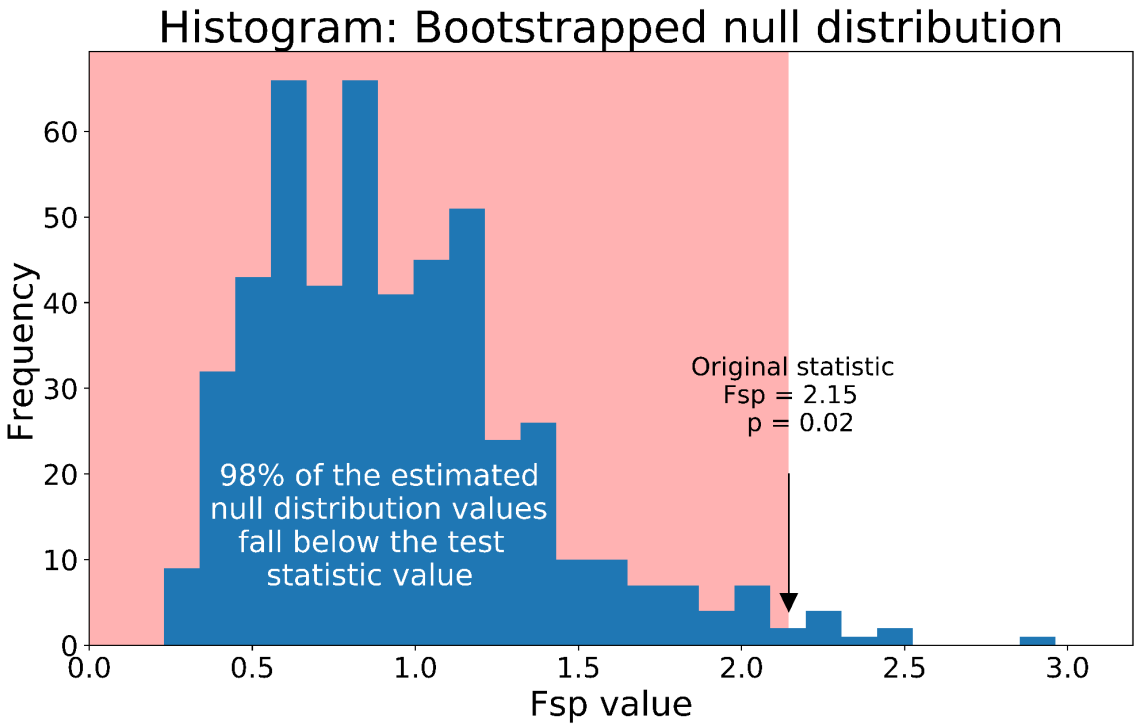

Figure 3-9 The estimated null distribution for a test statistic generated using the bootstrap. The significance of the original test statistic (from the coherently arranged ensemble data) can be obtained by evaluating its position within the estimated null distribution. The test statistic lies above 98% of the values of the estimated null distribution, providing a *p* value of 0.02, indicating that the null hypothesis of no ABR being present, can be rejected (for a significance level of 0.05). Figure redrawn with permission based on a figure in Chesnaye (2019).

3.3.1 Conclusion

Effective prediction of hearing thresholds using the ABR is in turn reliant upon effective detection of the response buried in background noise several orders of magnitude larger. Through collecting repeated recording epochs, coherent averaging may be used to improve the SNR of the measurement. Several statistical detection methods have been shown to be effective at detecting the ABR. These include the Fsp, the Fmp, the q-sample uniform scores test and its modified versions, as well as Hotelling's T^2 test. In the next chapter we shall review how machine learning has been applied effectively to numerous clinical signal detection challenges and how it may be used to detect the ABR.

Chapter 4 ABR Detection using Machine Learning

4.1 Introduction

As discussed in Chapter 2, the ABR is an important clinical tool used to objectively assess individuals' hearing thresholds. It is especially useful for assessing individuals for whom it may not be possible to obtain reliable behavioural hearing thresholds, e.g. newborn babies, infants, some adults with learning disabilities, and some adults with cognitive impairment (Section 2.1.4.2—Hearing Threshold Estimation). In clinical practice, ABR interpretation is based on visual inspection of the coherently averaged waveform by clinicians. Interpretation based on visual inspection is known to be variable, even amongst experienced clinicians (Cohen *et al.*, 1971; Vidler and Parker, 2004) (Section 3.2.2—Visual Inspection). Whilst several statistical ABR detection methods exist, visual inspection is still purported to be the gold standard and remains the basis for ABR detection in clinical practice (British Society of Audiology, 2019c).

The use of machine learning techniques is becoming increasingly prevalent in the biomedical literature as the effectiveness of these methods become more widely recognised and adopted. Machine learning algorithms have been found to exceed the contemporary gold-standard of human expert performance in a number of clinical fields (Sidey-Gibbons and Sidey-Gibbons, 2019). Examples of this include ECG interpretation (Hannun *et al.*, 2019), retinal disease detection from optical coherence tomography (OCT) scans (De Fauw *et al.*, 2018), lung cancer screening (Ardila *et al.*, 2019), and EEG interpretation (Medvedev, Agoureeva and Murro, 2019). Objective detection methods have the potential to assist clinicians in deciding whether an ABR response is present or absent. This has the potential to save clinical staff time by reducing the time required to interpret ABR waveforms. This has an associated cost benefit and frees up clinical staff time to be used for other endeavours. Automated ABR detection also has the potential to exceed the performance of human clinicians which, by extension, would be expected to lead to better patient outcomes. The purpose of the work presented in the current chapter is to leverage the effects of machine learning and apply these to the task of ABR detection. In this chapter the state of the literature regarding the detection of the ABR using machine learning will be discussed. Subsequently, a study will be presented whereby the performance of a trained machine learning algorithm is compared to that of prominent statistical ABR detection methods.

4.1.1 Chapter-Specific Acknowledgements

Please note that most of the findings of the study presented in this chapter have been published in McKearney *et al*. (2021). The figures from this published article are adapted with permission from Wolters Kluwer Health, Inc.: McKearney RM, Bell SL, Chesnaye MA, and Simpson DM. (2022) 'Auditory Brainstem Response Detection Using Machine Learning: A Comparison With Statistical Detection Methods', *Ear & Hearing*, 43(3), pp. 949–960, doi: 10.1097/AUD.0000000000001151. The written contents of this article are paraphrased from the final peer-reviewed manuscript for use in this thesis chapters as permitted by Kluwer Law International who provide the right to reproduce and distribute one's published work in order to further one's career, e.g. for use in non-commercial dissertations. Regarding author contributions, the published manuscript was written by the present author with feedback provided by all other authors (Prof. Steven Bell, Dr Michael Chesnaye, and Prof. David Simpson). This thesis chapter is based on the published article but has been rewritten and includes additional content. The code for the frequency domain bootstrap was provided by Dr Michael Chesnaye. Prof. Steven Bell and Prof. David Simpson (PhD supervisors of the present author) provided supervisory guidance regarding all aspects of the study.

4.1.2 Literature Review

The first study to use machine learning for ABR detection was published in 1991 by Alpsan (1991). In this study, 285 EEG recordings were labelled independently by three experts as belonging to one of two classes: 'Response' or 'No Response'. EEG waveforms which could not be clearly labelled by the experts were discarded and not used to train or evaluate the machine learning algorithm. The discarded waveforms were reported as being those recorded at low stimulus levels or containing noise, representing those waveforms most challenging to label. Discarding these waveforms therefore overly simplifies the task, biasing the performance level reported to being overly optimistic (McKearney and MacKinnon, 2019). Inter-observer agreement of the three labelling experts was 78.2%, with 21.8% of waveforms therefore being considered difficult to label and discarded. A three-layered artificial neural network (multilayer perceptron—MLP) was trained and tested on the averaged EEG waveforms which had been smoothed and compressed prior to normalisation or scaling (both of these were evaluated separately), achieving a maximum accuracy of 75%. A breakdown of the results is shown in Figure 4-1. It can be observed that there is a class imbalance with ~4:1 ratio of 'response present' to 'response absent' data. This may inform interpretation of the reported accuracy statistic, which is known to be an unreliable outcome measure for imbalanced datasets (Luque *et al.*, 2019).

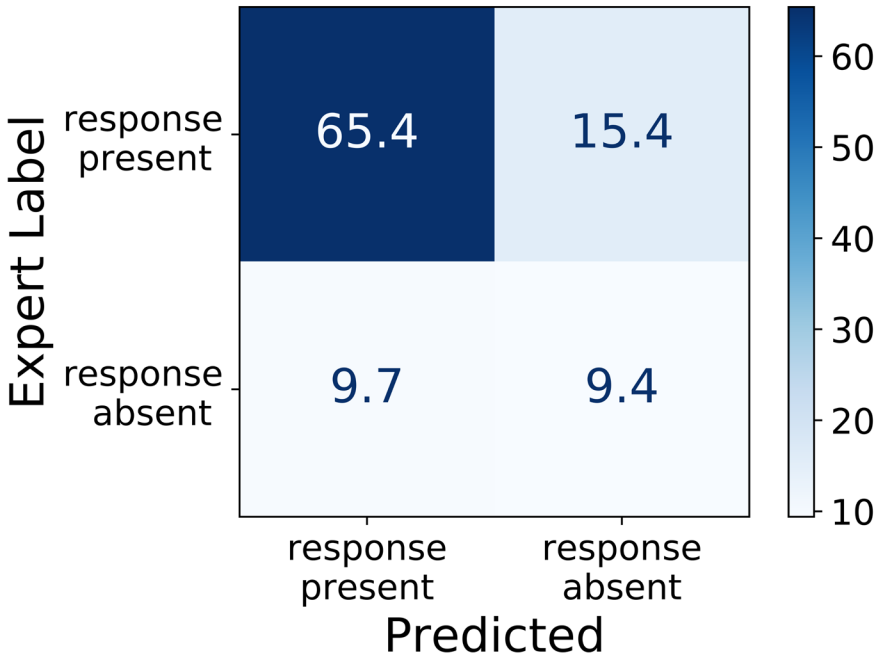

Figure 4-1 Confusion matrix showing the results from Alpsan (1991). The results represent the average percentage score obtained over 10 trials using randomly initialised neural network weights. The predicted results of the neural network were compared to the labels provided by three human experts. Note that the results do not sum to 100 percent, presumably as a result of rounding.

A very small amount of the data was used for training—up to 60 training instances from the dataset, with the remining data being used to evaluate the model performance. It is reported that test set performance saturated above a training set size of $n = 45$. This is surprising given the large amounts of data typically required to effectively train machine learning algorithms. This finding may reflect a restricted number of parameters within the model, giving it a low capacity to learn and therefore making it liable to underfitting. The hyperparameters (the number of units in the one hidden layer in this study) appear to have been optimised using the performance on the test set data which could result in an overly optimistic bias of the model's generalizable performance on unseen data. An extension to this work was published by Alpsan *et al.* (1994), with a focus on how hyperparameter optimisation of a multilayer perceptron affected detection performance. In summary, the authors found that hyperparameter optimisation was a worthwhile endeavour with some hyperparameter combinations noticeably outperforming others.

Acir, Özdamar and Güzeliş (2006) trained a support vector machine to classify EEG data as either ABR 'response present' or 'response absent'. Three separate feature sets were evaluated:

1. Amplitude values of the normalised coherent average.
2. Discrete cosine transform (DCT) coefficients.

3. Discrete wavelet transform (DWT) coefficients.

After feature extraction (applied to the normalised coherent average) a feature selection process was applied to select the most salient features for each of the three feature sets. This was performed by applying a sensitivity analysis (Belue and Bauer, 1995) which calculates the sensitivity of the classifier's output to changes to its input. The sensitivity analysis was performed iteratively, with the support vector machine (SVM) classifier being trained on increasingly reduced feature subsets from the training set and then evaluated on the test data over each iteration. A limitation of this study is therefore that feature selection appears to be informed by data from the test set, which likely serves to leak information from the test set into the choice of features, resulting in an overly optimistic bias in the reported test set results (McKearney and MacKinnon, 2019). The highest score reported using the test set data was 97.7% (n=180 test set) using the DCT coefficient feature set. Performance of the SVM was compared to labels provided by a human expert.

In 2007, Davey *et al.* reported the use of a hybrid model making use of both time and frequency domain features to classify ABR waveforms. The data used consisted of EEG waveforms recorded from 85 subjects of varying hearing status. The data were labelled as belonging into one of two classes ('response present' or 'response absent') by a human expert. A summary of the classification process is provided in Figure 4-2. For potential small responses with a mean pre-stimulus-to-mean post-stimulus power ratio of <5, a hybrid classification system was employed by combining the predictions of a time domain and a frequency domain classifier. An artificial neural network or a decision tree were used as time-domain and frequency domain classifiers and optimised to see which performed best. The two predictions were combined by using a Dempster-Shafer discounting factor (Shafer, 1976; Liu, 2001). Dempster-Shafer theory provides a method of combining evidence (in this case model predictions) from different sources in order to come to a more complete estimate regarding the hypothesis (Bezerra *et al.*, 2021). A validation data portion within each iteration of k-fold cross-validation was used to inform the value of the discounting factor. This study is one of the few studies in the field of ABR to apply k-fold cross-validation when evaluating model performance. For the data where the mean pre-stimulus-to-mean post-stimulus power ratio was >5 the hybrid classifier achieved an accuracy was 95.6%. For the potential small response data where the mean pre-stimulus-to-mean post-stimulus power ratio power ratio was <5, the hybrid algorithm classification accuracy was 85.0%. This reporting of the model performance highlights how difference performance can be depending on the quality of the data (McKearney and MacKinnon, 2019); it is much easier for a classifier to achieve a high accuracy if the SNR of the 'response present' data is high and if the noise level in the 'response absent' data is low.

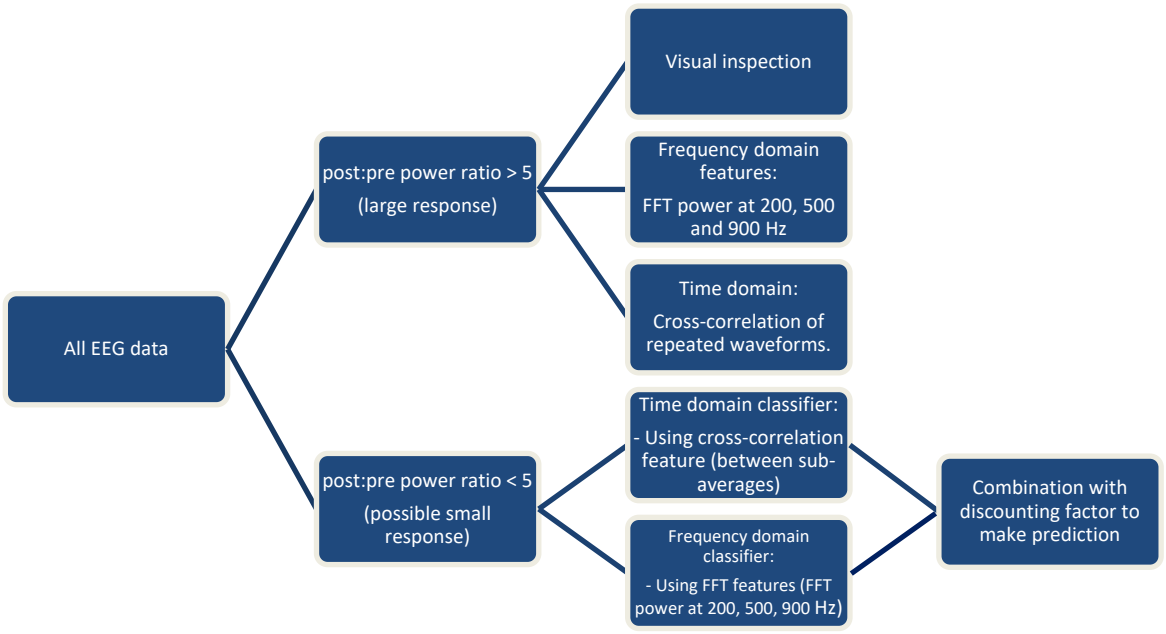

Figure 4-2 The outline of the hybrid model used by Davey *et al.* (2007) to classify EEG waveforms. The first stage is to differentiate large responses from potential small responses using the mean pre-stimulus-to-mean post-stimulus power ratio power ratio. For large responses where the power ratio was >5, one method of either: visual inspection, FFT power, or cross-correlation was used to confirm the presence of a response, ensuring that it was not artefact. For waveforms with a mean pre-stimulus-to-mean post-stimulus power ratio power ratio of <5, where a potential small response was present, the next step was to combine the predictions of a time-domain and a frequency-domain classifier using a discounting factor in order to make the final prediction. An artificial neural network or a decision tree were used as time-domain and frequency domain classifiers and optimised to see which performed best. Figure redrawn from Davey, R. *et al*. (2007) 'Auditory brainstem response classification: A hybrid model using time and frequency features', Artificial Intelligence in Medicine. Elsevier, 40(1), pp. 1–14. doi: 10.1016/J.ARTMED.2006.07.001., with permission from Elsevier.

Another study using wavelet analysis for feature extraction was reported by Rahbar *et al.* (2007). As the ABR latency is variable, being affected by stimulus level, a dual-tree complex wavelet transform was used as this technique is suggested to avoid shift variance. Shift invariance is theorised to improve ABR detection (Rahbar *et al.*, 2007). Data were divided into a training set and a test set. A three-layered multilayer perceptron (MLP) was used to classify the extracted features into one of three classes: 'response present with a normal wave V', 'response present with no clear wave V', or 'response absent'. The mean accuracy across all three classes was 83.3%.

In another study using DWT-extracted features combined with a MLP, Dass, Holi and Soundararajan (2016) used a combination of time and frequency domain features. An accuracy of 90.7% was achieved. Interestingly, the time domain features included the peak latencies of waves I, III, and V. It is not explained in the article how these were obtained and whether these were manually extracted by humans or automated. Neither is it clear what the wave latency input data values would be should the waveform be considered 'response absent' and therefore not exhibit any wave peaks. Following the theme of DWT features, Zhang *et al.* (2005) combine DWT feature extraction with a Bayesian network classifier, achieving an accuracy of 78.9%. A Bior 5.5 wavelet was used.

McCullagh *et al.* (2007) compared the performance of four different classification methods at detecting the ABR using 10-fold cross-validation: SVM, MLP, Naïve Bayes, and KStar algorithms. A clinical expert labelled the 550 EEG waveforms as being either 'response present' or 'response absent'. All four algorithms performed well (81.9–83.4% accuracy), with the highest accuracy being achieved using the Naïve Bayes algorithm.

Acir, Erkan and Bahtiyar (2013), used a support vector machine to detect the ABR, but with a focus on comparing two particular feature extraction methods:

1. DWT coefficients extracted from the unweighted coherent average.
2. DWT coefficients extracted from an estimated signal using a wavelet network-based adaptive estimation.

The highest accuracy achieved was 96.0%, which in required only 64 recording epochs per waveform to reach this level of performance. A limitation of this study is that test data appear to be used during the feature selection process to calculate the feature saliency values. This would leak information from the test set data into the features chosen, meaning that the test set score is likely to be overly optimistic in its estimation of how the model would perform on unseen data.

In a more recent study, McKearney and MacKinnon (2019) used k-fold cross-validation to compare the performance of a variety of deep learning algorithms before evaluating the best one on a separate test set. The data were labelled by two clinicians as belonging to one of three classes: 'response absent', 'inconclusive' or 'clear response'. These classes reflect the decision criteria used by the British Society of Audiology (2019c), acknowledging that some waveforms contain insufficient information to confidently determine whether a response is truly present or absent at the stimulus level in question. Information from 'inconclusive' waveforms would not be used to inform the decision of the level of the ABR threshold. The test set accuracy achieved across all three classes was 92.9%.

More recently still, Thalmeier *et al.* (2021) evaluated two approaches for classifying whether averaged raw EEG recorded from mice contained an ABR or not. The first approach was to use a supervised pair of convolutional neural networks. The first CNN would predict if an ABR response was present or absent for each stimulus level. The second CNN would then estimate the hearing threshold using the outputs of the first CNN as its input. The second approach evaluated was a self-supervised method termed by the authors as 'sound level regression'. Here a random forest regression model would first predict the sound stimulus level used to evoke a given response, for each of the recordings made (across stimulus levels). It seems initially counter-intuitive to predict the stimulus level, however, the authors clarify that this step is essentially used to produce a series of values for use in step 2, with the sound level predictions only being reliably possible if the actual stimulus level used is above that of the hearing threshold. A function was then fitted to the sound level predictions from step 1 and used to predict the hearing threshold. The supervised neural network (maximum 77.7% within ±5 dB SL of the human-defined label) performed better than the self-supervised 'sound level regression' method (maximum 72.1% within ±5 dB SL). However, the authors note that supervised algorithms require human experts to spend lots of time labelling the data. The self-supervised method is therefore proposed as an effective alternative which can be used readily on any ABR dataset without the need for data labelling (Thalmeier *et al.*, 2021).

4.1.3 Challenges in this Field

It is extremely challenging to draw any meaningful comparisons between the results of the various studies using machine learning to detect the ABR. This is a by-product of the substantial heterogeneity present between studies in terms of the data used, data labelling processes, methodology employed, and outcome measures reported (McKearney and MacKinnon, 2019).

Differences between datasets may account substantially for the differences in performance observed between detection algorithms. For example, detection will be undeniably easier for datasets containing high-SNR data (e.g. if a higher stimulus level is used) than with datasets containing low-SNR data. Of the results presented in Section 4.1.2, accuracies range from 74.8–96.0%. To be truly informative, the accuracy needs to be interpreted in light of the class balance and the characteristics of the dataset (e.g. the SNR of 'response present' data) and ideally be supplemented by clinically relevant outcome measures (such as sensitivity/specificity performance). Accuracy remains among the most popularly reported outcome measures in the studies presented in the Literature Review. However, this outcome measure becomes less meaningful if the data are imbalanced, e.g. if one class is underrepresented. In the case of ABR detection, additional outcome measures which reflect the nature of the test's clinical use are

beneficial. For example, having a high specificity is important as it is clinically desirable to have a low false positive rate in order to avoid falsely determining that an infant can hear normally when they in fact cannot (British Society of Audiology, 2019c). The differences in outcome measure used between studies again makes drawing meaningful comparisons between studies challenging. Additionally, the methodologies employed by the studies in this field vary significantly, with some studies more diligently than others employing methodologies such as cross-validation to obtain unbiased estimates of their model's generalisable performance.

At present, there is really no benchmark by which to compare performance to see if any newly proposed approach is better than any previous one. The use of a standardised published dataset may help to solve this problem (as used in Kaggle machine learning competitions—Kaggle, 2021), but would introduce its own limitations, such as perhaps limiting research to focus on data recorded using one particular set of recording parameters. There is also the potential for chance high performance findings to occur due to multiple algorithms being evaluated using the same dataset. It could however be argued that this is still the case should different datasets be used. One way of providing a performance benchmark, allowing comparison between studies, could be to compare machine learning algorithm performance to that of prominent statistical ABR detection methods (as presented in Section 3.2—ABR Detection Methods) (McKearney *et al.*, 2022). Dependant on the outcome measures reported, this would allow a degree of comparability between studies using heterogenous datasets. None of the previous studies in this field have compared the performance of their proposed algorithm with a statistical detection method.

Data labelling is an integral part of the supervised machine learning process, affecting both the training of the algorithm as well as determining which of the predictions of the algorithm are correct. Considering that this field is largely driven by the pursuit of providing an automated objective detection algorithm, it is somewhat paradoxical that the performance of these algorithms be compared to the yardstick of subjective human interpretation (McKearney *et al.*, 2022). A supervised machine learning algorithm may only be as effective as the subjective ratings of the human experts, with algorithmic predictions being incorrect by virtue of not matching those of the human experts—even if the experts are wrong. Using this subjective human-defined approach to labelling may allow machine learning algorithms to match the performance of human experts, which is helpful in standardising access to expert levels of signal interpretation which may not be readily accessible in regions where access to resources and clinical training may be limited. But why set the limit at human expert level performance when it is known that machine learning algorithms are able to surpass this level of performance in a number of different medical detection tasks (Haenssle *et al.*, 2018; Yim *et al.*, 2020). One method of partially overcoming the limitation of subjective human-defined labels is to standardise the recording process to label

recordings measured when no auditory stimulus is delivered to the ear as 'response absent' (e.g. using a clamped earphone tube, allowing the same environmental electromagnetic activity but preventing sound from arriving at the eardrum). Recordings obtained using a suprathreshold stimulus may be considered 'response present' with acknowledgment that near threshold labelling may become less accurate. Provided that certainty is present for the 'no response' class data, which should be achievable with appropriate experimental rigour, this will allow for effective training and evaluation of a machine learning algorithm. Alternatively, in order to definitively control the labels of both classes of data, simulation may be used whereby ABR templates are added to no-stimulus EEG data. The SNR of the data and true noise levels may therefore be known, allowing the ground truth ('response present'/'response absent') of the data to be known. Simulation additionally allows for large amounts of data to be produced, which is necessary to effectively train and evaluate data-hungry machine learning algorithms.

This study addresses some of the limitations in the field, as highlighted above, in order to evaluate various machine learning algorithms effectively and compare their performance to that of prominent statistical ABR detection methods.

4.1.4 Aims and Objectives

Aim 1. Develop a suitable database of 'response present' and 'response absent' data by which to train and evaluate machine learning algorithms.

Objective 1a: Use simulation to boost the amount of training data available.

Objective 1b: Add ABR templates to half of the data in order to control the ground truth labels.

Aim 2. To train a machine learning algorithm to be able to determine whether EEG data contains an ABR response or not.

Objective 2a: Compare several prominent machine learning approaches using nested k-fold cross-validation in order to select the best approach.

Objective 2b: Select the best machine learning algorithm for evaluation on the separate, previously unseen, test set.

Aim 3. Compare the performance of the machine learning algorithm with that of prominent statistical ABR detection methods.

Objective 3a: Prominent statistical ABR detection methods to evaluate are the Fsp, Fmp, Hotelling's T^2 test, and the q-sample uniform scores test.

Objective 2b: Evaluate sensitivity and specificity performance on the test set data with reference to the known ground truth labels.

4.2 Methods

4.2.1 Data

4.2.1.1 Subject Recorded ABR Data

The ABR data used in this study have been described in previously published works (Lv, Simpson and Bell, 2007; Chesnaye *et al.*, 2018; Chesnaye, 2019), and were made available by Dr Michael Chesnaye in the University of Southampton Institutional Repository (doi:10.5258/SOTON/D0168). The ABR data were recorded from 12 participants (six female; six male), aged 18–30 years, with normal hearing (audiometric thresholds ≤ 20 dB HL, tested at octave intervals between 250 and 8,000 Hz) (Lv, Simpson and Bell, 2007). The electrode montage comprised of the noninverting electrode being placed in the vertex (Cz) position, the inverting electrode being sited on the nape of the neck, and the common electrode being placed in the frontal (F_Z) position. Electrode impedances were monitored to be below 5 kΩ throughout the recordings. A 100-μs click stimulus was delivered via ER-2 insert earphones (Etymotic, USA), at a stimulus rate of 33.3 Hz. Recordings were made at stimulus levels between 0–50 dB SL (sensation level), in 10-dB increments (Lv, Simpson and Bell, 2007). The click sensation level threshold (0 dB SL) was obtained using a 10-dB-down, 5-dB-up procedure (Lv, Simpson and Bell, 2007). A recording window of 30.03 ms was used (following the delivery of each stimulus), with the signal sampled at a rate of 10 kHz.

Offline processing of the recorded signals consisted of band-pass filtering from 30 to 1,500 Hz, using a 3rd-order Butterworth filter, and downsampling the signal from 10 kHz to 5 kHz. The filter settings used reflect the recommendations made by the British Society of Audiology (2019c).

For the current study, artefact rejection was additionally applied offline, with the rejection level set at ±15 μV. There were ~3,400 epochs on average in each recording, after artefact rejection was applied.

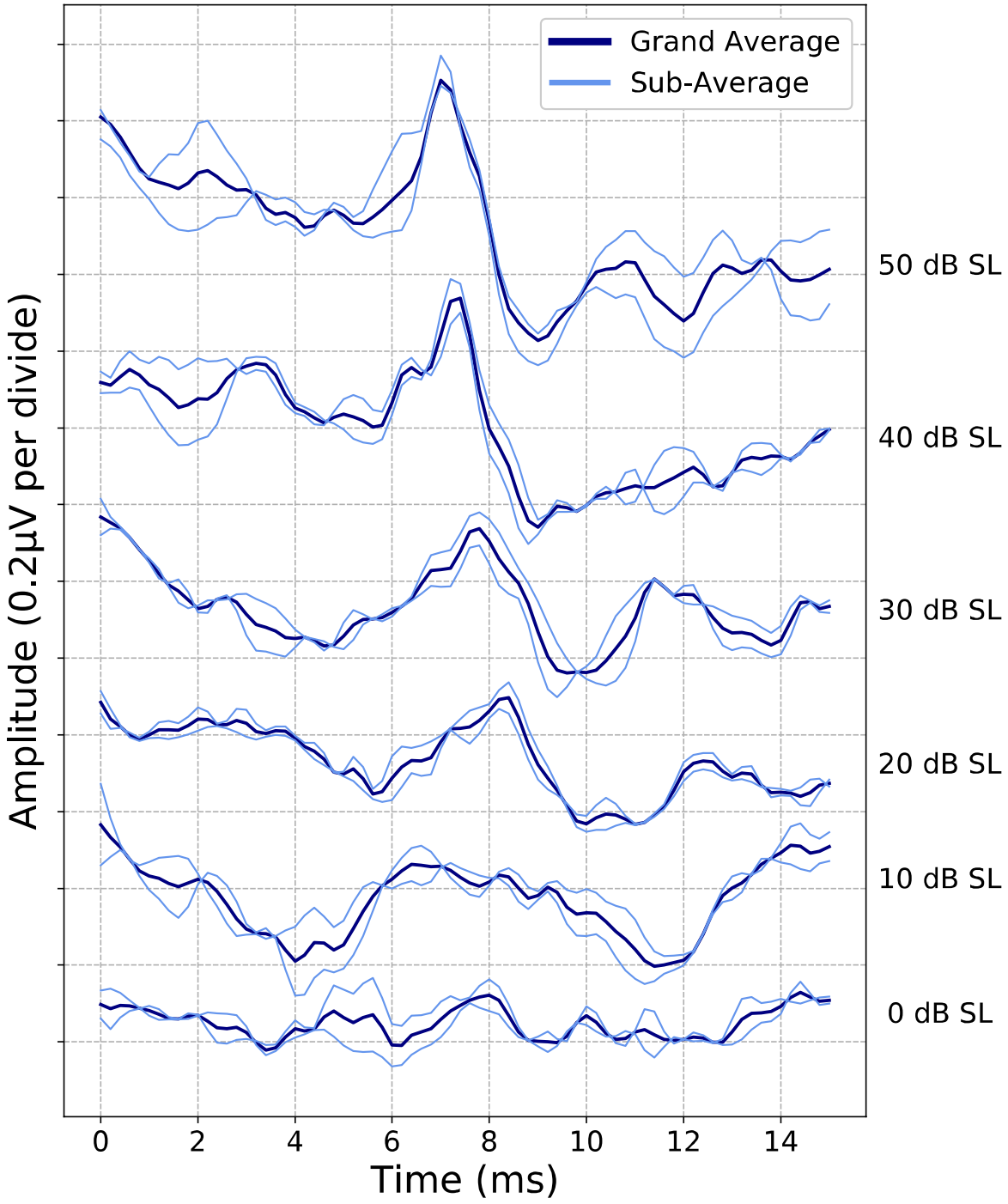

Figure 4-3 ABR waveforms. This figure shows example ABR data recorded from one participant across a range of stimulus levels (0 to 50 dB SL, in 10-dB increments). The bold line is the grand average of the two sub-averages at each stimulus level. Adapted with permission from Wolters Kluwer Health, Inc.: McKearney RM, Bell SL, Chesnaye MA, and Simpson DM. (2022) 'Auditory Brainstem Response Detection Using Machine Learning: A Comparison With Statistical Detection Methods', *Ear & Hearing*, 43(3), pp. 949–960, doi: 10.1097/AUD.0000000000001151.

4.2.1.2 Subject Recorded No-stimulus EEG Data

The subject recorded no-stimulus EEG data have been previously described in the literature (Madsen, 2010; Chesnaye *et al.*, 2018; Madsen *et al.*, 2018; Chesnaye, 2019; McKearney *et al.*, 2022). These data were recorded from 17 participants (5 female; 12 male), aged 24–52 years. EEG recordings were made under four separate recording conditions; asleep, lying still, blink (where subjects were prompted to blink every 1–3 seconds), and movement (where participants were instructed to move their heads) (Figure 4-4). Recordings were made using a Compumedics SynAmps 2 EEG amplifier, in an electrically shielded and acoustically isolated booth. The electrode montage consisted of a noninverting electrode sited in the left mastoid position, an inverting electrode sited on high forehead (F_Z), and a common electrode placed on the right cheek. The sampling rate used was 20 kHz (Madsen, 2010; Madsen *et al.*, 2018).

The no-stimulus EEG data were pre-processed offline using the same procedure as that applied to the ABR data; the data were downsampled from 20 kHz to 5 kHz, band-pass filtered from 30 to 1,500 Hz using a 3rd-order Butterworth filter, and artefact rejection was applied (±15 µV).

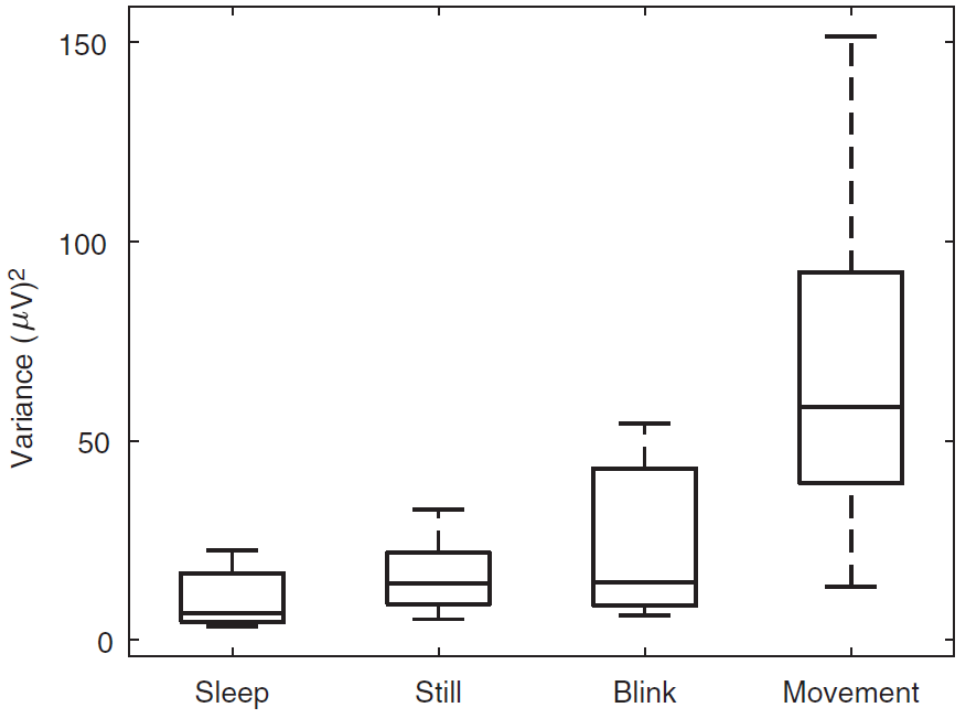

Figure 4-4 Characteristics of the no-stimulus EEG database. The data were recorded under four conditions ('sleep', 'still', 'blink' and 'movement'). The noise level within each group of recordings was quantified by the variance of the recordings. Note that variance was measured from the raw EEG data before artefact rejection had been applied. This figure is reproduced from: Madsen, S. M. K. *et al.* (2018) 'Accuracy of averaged auditory brainstem response amplitude and latency estimates', International Journal

4.2.2 Ethics

Overarching ethical approval was granted by the University of Southampton Faculty Ethics Committee to use the data from the subject recorded ABR and no-stimulus EEG datasets for the purpose of secondary data analysis for a range of research activities throughout these PhD studies (ERGO 55576).

4.2.3 Data Partitioning

Machine learning algorithms typically require large amounts of data to learn how to perform a task. The amount of data required will depend on the nature and difficulty of the task as well as the characteristics of the data. In order to have sufficient data, data were simulated using the subject recorded ABR data and the no-stimulus EEG data. The frequency domain bootstrap (FDB) (Paparoditis, 2002; Chesnaye *et al.*, 2021) was used to generate a large number of realistic no-stimulus EEG ensembles, based on the characteristics of the no-stimulus EEG database. I am grateful to Dr Michael Chesnaye who wrote the code for the frequency domain bootstrap which was used for the study presented in this chapter. The FDB is a parametric bootstrap technique whereby a section of EEG can be used to generate numerous surrogate EEG portions, which each reflect the spectral composition and change in amplitude over time of the original EEG sample (Figure 4-5) (Paparoditis, 2002; Chesnaye *et al.*, 2021).

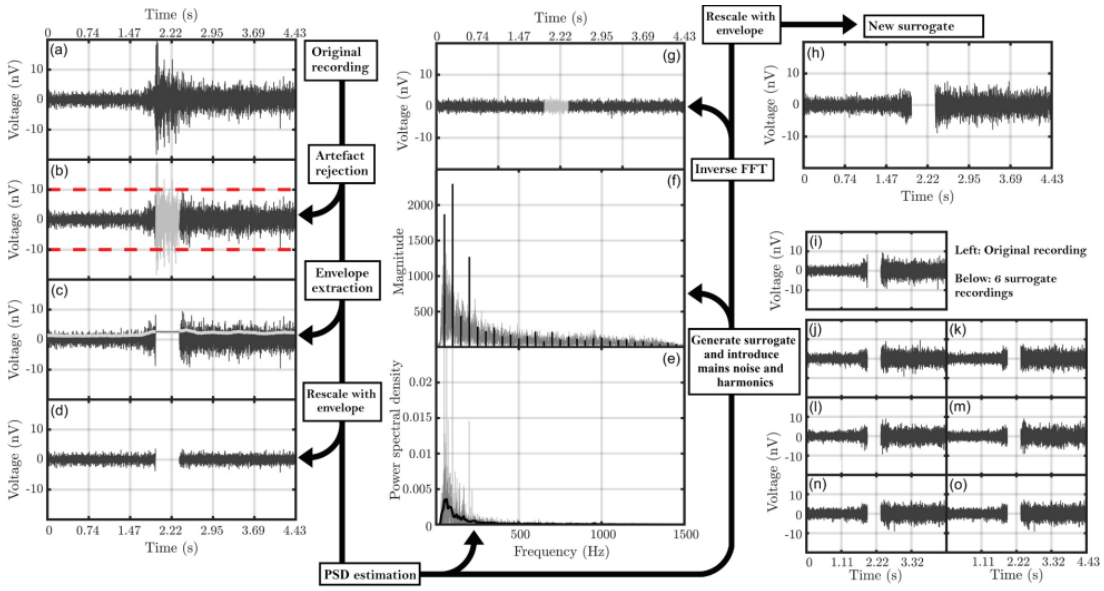

Figure 4-5 The frequency domain bootstrap procedure. The original EEG recording (a) is pre-processed as desired (b), prior to the envelope being extracted (c) and used to rescale the EEG (d). The power spectral density (PSD) of the rescaled EEG is then estimated and used to generate random PSD surrogates which are converted into magnitudes (f). An inverse FFT is then applied (g) before rescaling the surrogate using the previously extracted envelope (h). The original recording (i) may be used to generate multiple realistic surrogates (j,k,l,m,n,o) (Chesnaye *et al.*, 2021). Figure reused from Chesnaye *et al.* (2021) with permission from Elsevier.

For the current study, EEG data from the 'sleep' and 'still' conditions only were used (representing 15 participants), as these conditions most accurately reflect those observed in the clinical setting when recording the ABR. In total, 15,000 ensembles (each comprising 1,000 recording epochs), were generated using the FDB. These data were partitioned into a training set, a threshold set (used to set the detection criterion for the machine learning algorithm), and a test set, with no participant overlap between the sets (Raschka, 2020). It is important to avoid participant overlap in order to avoid overfitting and an overly optimistic assessment of an algorithm's generalisable performance.

For half of the data in the training set and the test set, an ABR template was added to the no-stimulus FDB-generated ensembles in order to simulate 'response present' data, with the template scaled such that the SNR of the simulated 'response present' ensemble matched that of the estimated SNR of the original subject recorded ABR ensemble used to generate the ABR template (McKearney *et al.*, 2022). The ensembles without an ABR template formed the 'response absent' data. In order to have high confidence that the ABR templates used contained a response, two audiologists were asked to independently assess the ABR waveforms, presented in the same

format as that displayed in Figure 4-3, and to label waveforms where a 'clear response' was deemed to be present, using the British Society of Audiology (BSA) criteria (2019c). Those coherent averages where both clinicians independently deemed there to be a 'clear response' were used as ABR templates to simulate the 'response present' data. Inter-observer agreement between the two audiologists was 93.1% (Cohen's kappa = 0.83). In order to avoid participant overlap between the training and the test set, the ABR participants were split between the training and the test sets. The threshold set (used to set the detection criterion for the machine learning algorithm) contained only 'response absent' data as 'response present' data are not required to set the detection criterion required to meet a target specificity level.

The ensembles in each set were split into 10 smaller constituent ensembles of varying size, including 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1,000 recording epochs. The training, threshold, and test set were made up of 90,000, 15,000, and 30,000 ensembles respectively (Figure 4-6).

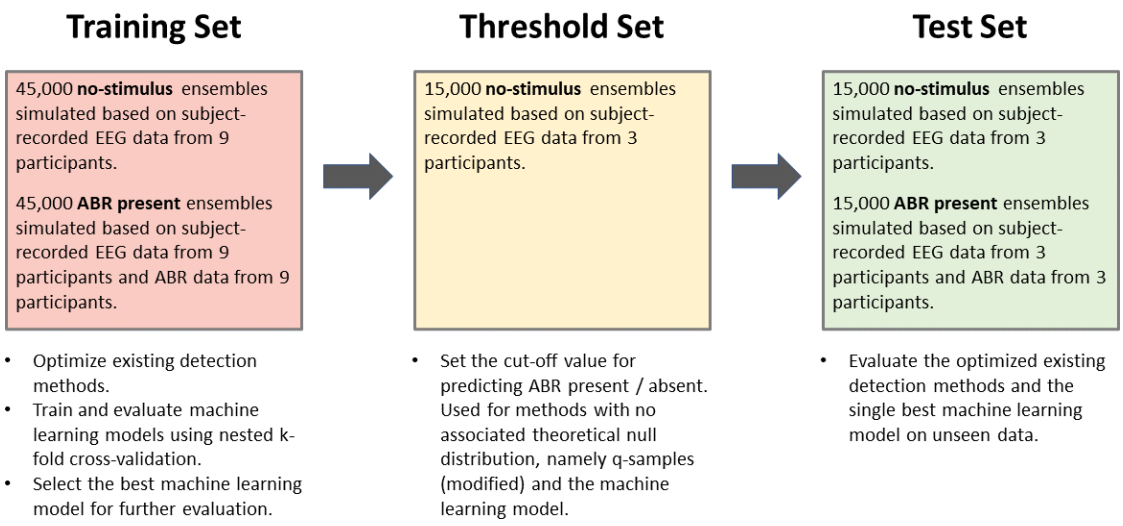

Figure 4-6 Data partitioning. The data were split into a training, threshold and a test set. There was no participant overlap between sets. In each set, there was an even split of ensembles of 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1,000 recording epochs. Adapted with permission from Wolters Kluwer Health, Inc.: McKearney RM, Bell SL, Chesnaye MA, and Simpson DM. (2022) 'Auditory Brainstem Response Detection Using Machine Learning: A Comparison With Statistical Detection Methods', *Ear & Hearing*, 43(3), pp. 949–960, doi: 10.1097/AUD.0000000000001151.

The training set was used to train and compare the performance of the proposed machine learning algorithms. The best machine learning algorithm, as determined from the training set data, was selected for evaluation on the previously unseen test set data. Prior to test set evaluation, the critical value for rejecting the null hypothesis ('response absent') needed to be

determined using the threshold set data. Having a critical value is necessary in order to be able to evaluate specificity performance using the test set data. These three steps will be described in more detail in the upcoming sections.

4.2.4 Nested K-Fold Cross-Validation on the Training Set Data

The training set data were used for multiple purposes, including selecting the best hyperparameters, training the different models, and selecting the best machine learning algorithm for further analysis on the test set. The 'no free lunch theorem' (Wolpert and Macready, 1997 in McKearney *et al.*, 2021) states that there is no single best machine learning approach to solve all problems; it is therefore necessary to consider several potential strategies. Several machine learning approaches were therefore considered. An effective way of combining these tasks in an unbiased manner is to use nested k-fold cross-validation (Varma and Simon, 2006; Bergstra and Bengio, 2012; Raschka, 2020). Here an inner loop of cross-validation is performed, nested within an outer loop of cross-validation (Figure 4-7). For each outer loop iteration, the inner loop of cross-validation is used to select the best hyperparameters, here defined as the highest mean area under the receiver operating characteristic curve (ROC AUC) score. The ROC curve provides a measure of a binary classifier's ability to discriminate between two different classes, as the threshold of the classifier is varied. The ROC AUC is the area under this curve, providing a single outcome measure, which is widely used to evaluate detection method performance (Fawcett, 2006). After the inner loop cross-validation, the model is then trained on the entire outer loop training fold data using the optimised hyperparameter combination. The trained model is then evaluated on the outer validation fold. This process is repeated in turn for each of the nine outer loop iterations, allowing a mean ROC AUC score across the outer validation folds to be calculated. The machine learning algorithm which achieved the highest mean ROC AUC score across the nine outer validation folds was selected as the best and used for subsequent analysis on the test set data. Each validation fold (for both the inner and outer loops) was contributed to by data from only one of the 15 no-stimulus EEG dataset participants, and only one of the 15 ABR dataset participants. There was therefore no participant overlap between training and validation folds during cross-validation (McKearney *et al.*, 2022). Leave-one-group-out cross-validation was used, with one group of data (10,000 ensembles) being used as the validation fold, both for the inner and outer loops. This meant that the outer loop contained nine groups, whilst the inner loop contained eight groups.

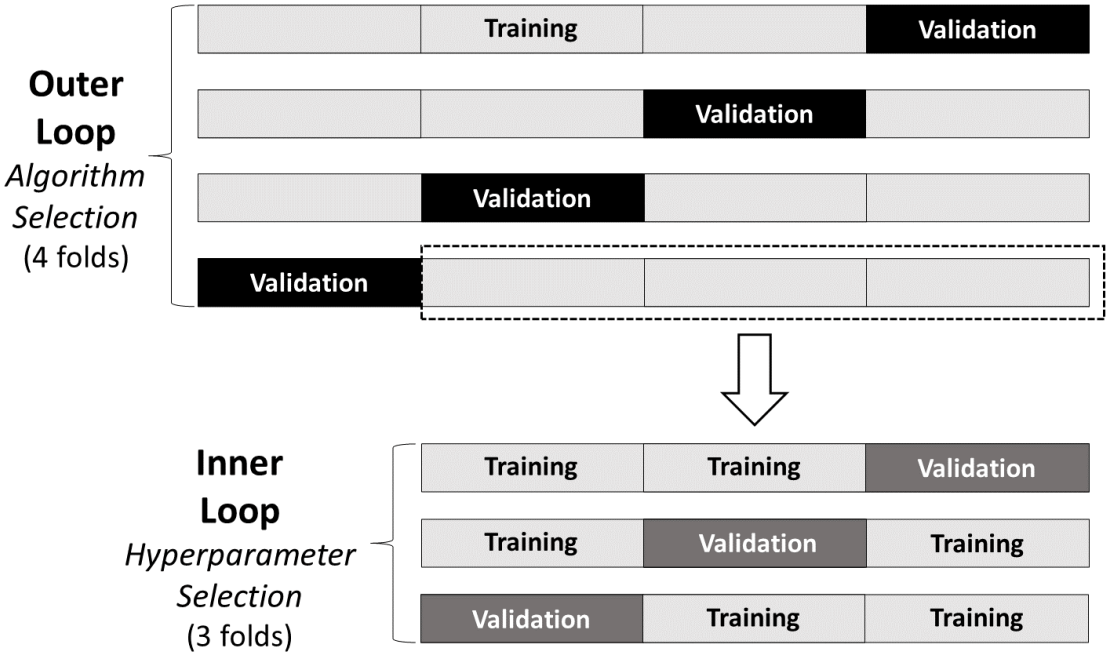

Figure 4-7 An illustration of nested k-fold cross-validation. For each of the four outer loop iterations, an inner loop of cross-validation is performed on the outer loop training fold in order to select the best hyperparameter combination. Following this, the model is trained on the entire outer loop training fold using the best hyperparameters, before being evaluated on the outer validation fold. The mean score across the four outer validation folds is used to select the best algorithm. Note that for simplicity, this figure represents a reduced version of the cross-validation procedure used; the study used nine outer loop iterations and eight inner loop iterations. Figure adapted from Raschka (2020) and Rashcka and Mirjalili (2017) with permission from Dr Sebastian Raschka and Packt (www.packtpub.com).

After completing nested k-fold cross-validation on the training set, the best machine learning algorithm was selected based on having the highest mean ROC AUC across the nine outer validation folds. The inner loop procedure of cross-validation was then applied to the entire training set data to select the best hyperparameter combination for the selected machine learning algorithm. Finally, the selected algorithm was trained using this hyperparameter combination on the entire training set.

4.2.5 Setting the Critical Value using the Threshold Set

Once the best machine learning algorithm had been selected and trained on the entire training set, using ROC AUC as the criterion for optimisation, it was necessary to select the critical value

for the detection method to determine if the null hypothesis of no response being present should be rejected. Unlike certain statistical detection methods such as the Fsp (Elberling and Don, 1984), the output of the machine learning algorithm does not follow a known theoretical distribution (McKearney *et al.*, 2022). One way of deriving the critical value is to obtain it empirically from a separate set of 'response absent' data—the threshold set. The desired false positive rate (FPR) chosen for this study was 1%. This relatively low FPR was selected, because in clinical practice audiologists require a high degree of certainty when declaring a response to be present in order to avoid mistakenly diagnosing a patient's hearing as being better than it truly is. This could potentially lead to patients not getting the hearing habilitation/rehabilitation that they require and lead to detrimental effects on, for example, a child's speech development (Kennedy *et al.*, 2006; Fulcher *et al.*, 2012).

In order to obtain the critical value that corresponds to a FPR of 1%, the trained machine learning model made a prediction for all of the threshold set data. This provided an estimated null distribution. The critical value was then taken as the value at which 99% of predictions for the 'response absent' data fell below. This process was performed separately for each of the separate ensemble sizes evaluated (100 up to 1,000 epochs).

As this method was not found to be effective at controlling the false positive rate across ensemble sizes, a second method for obtaining the critical value was also used which did not use the threshold set data—the bootstrap method (Section 3.3). The bootstrap method allows the critical value for the test statistic to be calculated for each individual ensemble being analysed. This obviates the need for a separate set of data from which to obtain the critical value (i.e. the threshold set) (Chesnaye *et al.*, 2018), and also allows the critical value to reflect the individual characteristics of the data being analysed.

4.2.6 Final Evaluation on the Test Set

Whilst the best machine learning algorithm was selected based on the highest ROC AUC score on the test set data, the ROC AUC metric does not provide a separate evaluation of the sensitivity and specificity performance of a classifier. It is important to know if a detection method is able to achieve the desired FPR to ensure that patients are not mistakenly diagnosed as being able to hear at a certain stimulus level when in fact they can't. It is also important to analyse the detection rate across ensemble sizes. A separate analysis of specificity and sensitivity was therefore conducted using the test set.

4.2.7 Statistical Detection Methods Evaluated

One of the aims of this study was to compare the performance of the machine learning algorithm with that of prominent statistical ABR detection methods. The statistical detection methods to which compare performance to include the Fsp (Section 3.2.3.1), the Fmp (Section 3.2.3.1), Hotelling's T^2 test (Section 3.2.3.2), the q-sample uniform scores test (Section 3.2.3.3), and the modified q-sample uniform scores test (modified version 2) (Section 3.2.3.4). Version 2 of the modified q-sample uniform scores test, which utilises the ranked phase angles in addition to the ranked spectral amplitudes, was chosen as it was found by Chesnaye *et al.* (2018) to have a higher detection rate compared to version 4 when using simulated data (Figure 3-6). All of these statistical detection methods, aside from the modified q-sample uniform scores test, produce test statistics which follow a known theoretical null distribution. The critical value of these detection methods may therefore be taken as the test statistic level which corresponds to a *p* value of <0.01 (the desired FPR). For the Fsp and the Fmp, the *p* values were obtained based on an *F*-distribution with $v_1 = 5$, and $v_2 = N - 1$ degrees of freedom (*df*) (Elberling and Don, 1984). For the modified q-sample uniform scores test, the critical value for each ensemble size was obtained empirically using the threshold set data (Section 4.2.5).

Some of these statistical detection methods have parameters which may be optimised to improve detection performance. These include the number of voltage means for the Hotelling's T^2 test, the number and range of spectral bands to be included for both the modified and unmodified version of the q-sample uniform scores test[1], as well as whether or not to use zero-padding to improve the FFT resolution, again for both the modified and unmodified versions of the q-sample uniform scores test. These parameters were optimised based on the mean ROC AUC score obtained using k-fold cross-validation performed on the training set data.

4.2.8 Machine Learning Approaches Evaluated

4.2.8.1 Multilayer Perceptron

A perceptron is a single layer of artificial neurones, whereby each separate neurone is fully connected to every input (Géron, 2017). A multilayer perceptron (MLP) is therefore several layers of perceptrons with each layer being fully connected with each previous layer. This network structure of multiple layers of inter-connected neurones is said to mimic the structure of the brain

[1] The author is grateful to Dr Michael Chesnaye for providing MATLAB code for the q-sample uniform scores test and its modifications, which was helpful when writing the functions for these equations in Python. The author is also thankful to Dr Michael Chesnaye for cross-checking the output of a Python implementation of Hotelling's T^2 test with that of a MATLAB implementation.

(McCulloch and Pitts, 1943). For each artificial neurone in the network, all of the inputs will be weighted and summed, before being passed to an activation function which determines the output of the artificial neurone (Géron, 2017). During training, artificial neurones in the MLP will form 'connections' (the weights) with other artificial neurones in the network. 'Connections' which improve performance (decrease the error) will be reinforced (increasing weight values), whereas 'connections' which degrade performance (increase the error) will be diminished (Géron, 2017).

Feature engineering (extracting meaningful features from the raw data) was used to extract features from each ensemble to be used as input variables to the MLP. First the coherent average of the ensemble was calculated. Secondly, the DWT was applied to the coherent average with three levels of decomposition, performed using a biorthogonal ('bior5.5') wavelet (Bradley and Wilson, 2004; Zhang *et al.*, 2004). Dimensionality reduction was achieved by extracting statistical features from each of the DWT coefficient subbands. These statistical features were the mean, mean of the absolute values, median, standard deviation, skew, kurtosis, RMS, variance, interquartile range, number of zero crossings, and the ratio of the mean absolute values between adjacent subbands. This feature extraction approach was based on that used by Subasi (2007) when detecting seizure activity from EEG data and Kandaswamy *et al.*, (2004).

The architecture of the multilayer perceptron is shown in Table 4-1.

Table 4-1 The architecture of the multilayer perceptron. Optimised hyperparameters are shown in *italics* and underlined. Hyperparameters which were not fine-tuned are shown in regular typeface.

Layer	**Hyperparameter settings**
Dense	43 units, input features=43, selu activation function
Dense	*Number of units*, selu activation function
Dropout	*Dropout rate*
Dense	*Number of units*, selu activation function
Dropout	*Dropout rate*
Dense	8 units, selu activation function
Dropout	*Dropout rate*
Dense	1 unit, sigmoid activation function

Hyperparameter optimisation was performed using a random search (using 90 hyperparameter combinations) of the hyperparameter space shown in Table 4-2.

Table 4-2 The hyperparameter space searched for the multilayer perceptron. Note that some hyperparameters were not searched—for these hyperparameters only one value is shown.

Hyperparameter	Values searched
Dropout rate	0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45
Number of training epochs	5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55
Learning rate	0.0004, 0.0006, 0.0008, 0.001, 0.0012, 0.0014
Number of units	20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70
Batch size	256
Optimiser	Adam
Loss function	Binary cross-entropy

4.2.8.2 Convolutional-LSTM

A convolutional long short-term memory network (CNN-LSTM) is constructed by combining a convolutional neural network (CNN) with one or more recurrent long short-term memory (LSTM) layers. Convolutional neural networks are a type of neural network which are commonly used for computer vision tasks (Chollet, 2018). CNNs are characterised by convolutional layers (LeCun *et al.*, 1998), often used in conjunction with pooling layers. Convolutional layers apply a convolution operation to the input, producing a 'feature map' (Chollet, 2018). The convolutional layer may be thought of as a filter whose coefficients (weights) are learnt from the training data. Pooling layers are used to downsample the feature maps produced by convolutional layers. Convolutional layers and pooling layers can work in combination to provide a degree of local translation invariance; this means that a response occurring at a different latency to that seen previously (during training) should still be interpreted as a response. Stacking convolutional layers allows machine learning algorithms to learn the temporal hierarchy of features (equivalent to spatial hierarchy in image processing) (Yamashita *et al.*, 2018 in McKearney *et al.*, 2022).

Recurrent neural networks (RNNs) are well suited to analysing sequence data such as EEG. Unlike feedforward neural networks (such as MLPs) where the activations flow only between the input and output layers, RNNs also utilise connections going backwards in time (Géron, 2017). This property imbues recurrent neurones with a 'memory' as its output at any given point in time is a function of all previous time steps (Géron, 2017).

The architecture of the CNN-LSTM used in the current study was two repeated one-dimensional convolutional/max pooling layers, followed by an LSTM layer and three fully connected layers (Table 4-3). As input, this model received three vector features: the coherent average, a denoised version of the coherent average using the Teager-Kaiser energy operator (TKEO) (Kaiser, 1990), and a vector of the *p* values for the Student's *t*-statistic as calculated down each column of the ensemble matrix ($\mathbf{X}$) (McKearney *et al.*, 2022).

Table 4-3 The convolutional long short-term memory network architecture. Optimised hyperparameters are shown in *italics* and underlined. Hyperparameters which were not fine-tuned are shown in regular typeface.

Layer	Hyperparameter settings
Convolutional 1D	relu activation function, *number of units*
Max Pooling	pool size=2, padding='same'
Convolutional 1D	relu activation function, *number of units*
Max Pooling	pool size=2, padding='same'
LSTM	*Dropout rate*, *recurrent dropout rate*, *number of units*
Dense	relu activation function, *number of units*
Dropout	*Dropout rate*
Dense	7 units, relu activation function
Dropout	*Dropout rate*
Dense	1 unit, sigmoid activation function

Hyperparameter optimisation was performed using a random search (using 90 hyperparameter combinations) of the hyperparameter space shown in Table 4-4.

Table 4-4 The hyperparameter space searched for the CNN-LSTM. Note that relatively low values for the number of training epochs were searched as the training set size was

quite large and to limit the computational expense when evaluating the model within nested the cross-validation procedure.

Hyperparameter	Values searched
Dropout rate	0, 0.05, 0.1, 0.15, 0.2
Recurrent dropout rate	0, 0.05, 0.1, 0.15, 0.2
Learning rate	Fifteen evenly spaced values in a log space between 0.005 and 0.25
Batch size	128, 256, 512, 1024
Convolutional kernel size	3, 5, 7
Training epochs	3, 4, 5, 6, 7, 8, 9, 10
Number of units	15, 18, 21, 24, 27, 30
Optimiser	Adam, stochastic gradient descent
Loss function	Binary cross-entropy

4.2.8.3 Random Forest

Random forests are made up of an ensemble of decision trees (Ho, 1995). Decision trees are machine learning algorithms which are able to learn to split data based on feature values using a decision algorithm and can be used for both classification and regression tasks (Raschka and Mirjalili, 2017). Individual decision trees are combined to form a random forest, which is an example of ensemble learning whereby multiple algorithms are combined to make a final classification/regression decision. Random forests are typically trained using 'bagging', a contraction of the words 'bootstrap' and 'aggregation' (Breiman, 1996 in McKearney *et al.*, 2022). Here, random forests are trained on a randomly selected subsection of the data (with replacement—i.e. bootstrapping) before their individual predictions are aggregated. This process is considered to reduce the variance of the model, improving its generalisable performance (Breiman, 1996 in McKearney *et al.*, 2022).

This model aimed to combine, using machine learning, the properties of the prominent ABR statistical detection methods. This was accomplished by extracting features from the raw EEG data, using these statistical detection methods. Specifically, the features were extracted using the Fmp, Hotelling's T^2 test (applied in 34 iterations each using a different TVM parameter from two

to 35 TVMs), q-sample uniform scores test, and the residual noise within the coherent average. These 37 features served as the input to the random forest. The *p* values were used rather than the raw test statistics produced by the statistical detection methods, because unlike the raw test statistics, the *p* values are not dependant on the ensemble size for their interpretation.

Hyperparameter optimisation was performed using a random search (using 90 hyperparameter combinations) of the hyperparameter space shown in Table 4-5.

Table 4-5 The hyperparameter space searched for the random forest. Note that some hyperparameters were not searched—for these hyperparameters only one value is shown.

Hyperparameter	Values searched
Maximum number of samples used to train each base estimator	2500, 3500, 4500, 5500, 6500, 7500
Number of trees in the random forest	5000
Maximum depth	10, 40, 70, 100, 130, 160, 190, None
Number of features considered when determining the best split	The square root of the number of features
Minimum number of samples per split	5, 10, 15
Minimum number of samples per leaf	1, 2
Criterion for measuring the quality of each split	gini, entropy

4.2.8.4 Stacked Ensemble

Like random forests, stacked ensembles make use of ensemble learning. Stacked ensembles combine the outputs of two or more base estimators using a meta-estimator which receives these base predictions as inputs and uses them in turn to make the final classification/regression decision (Wolpert, 1992). Combining multiple base estimators, which may each consider the data

in different ways, can improve generalisable performance. The two base estimators used in the stacked ensemble were the previously discussed convolutional-LSTM (Section 4.2.8.2) and random forest (Section 4.2.8.3). This approach was chosen in order to combine the template-matching approach of the convolutional-LSTM with the random forest's combined statistical detection approach, reducing over-reliance on any one input feature/approach. For each ensemble being evaluated, each of the two base estimators would produce an output prediction which in turn acted as the input variables to a logistic regression classifier meta-estimator in order to make a final single prediction. Hyperparameter optimisation was performed using a random search (using 90 hyperparameter combinations) of the combined hyperparameter space already displayed for the CNN-LSTM (Table 4-3, Table 4-4) and for the random forest (Table 4-5) (the two base estimators in the stacked ensemble) as well as the regularisation hyperparameter C for the logistic regression meta-estimator (Table 4-6). An illustration of the stacked ensemble architecture is provided in Appendix A.

Table 4-6 The hyperparameter space searched for the logistic regression meta-estimator.

Hyperparameter	Values searched
Regularisation hyperparameter C	Eleven evenly spaced values on a log scale between 1e-4 and 10,000

The machine learning algorithms compared also reflect the choice of input features used and so cannot be considered in isolation.

The machine learning algorithms were constructed using the scikit-learn (Pedregosa *et al.*, 2011) and Keras (Chollet and and others, 2015) Python software libraries.

4.3 Results

4.3.1 Optimisation of the Statistical Detection Methods

Cross-validation using the training set data was used to optimise the Hotelling's T^2 test, as well as the original and modified versions of the q-sample uniform scores test. For Hotelling's T^2 test, the optimal number of voltage means was 16 (Figure 4-8), with a stable and high level of performance observed between 15–30 voltage means.

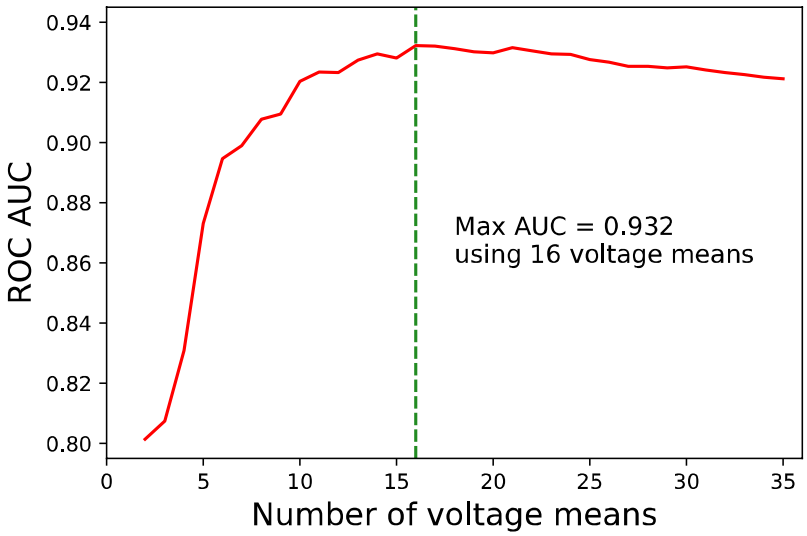

Figure 4-8 Optimisation of the number of voltage means used in the Hotelling's T^2 test.

Both q-sample uniform scores test methods were found to achieve higher ROC AUC score when zero-padding was applied. The optimised original q-sample uniform scores test used 150 spectral bands in the range 30–600 Hz. The optimised modified q-sample uniform scores test used 150 spectral bands in the range 30–1350 Hz.

4.3.2 Training Set Cross-Validation

The ROC AUC scores across the nine outer loop validation folds were compared across all of the detection methods evaluated (both traditional statistical and machine learning detection methods). The ROC AUC scores for each detection method were visually inspected using Q-Q plots and found not to be normally distributed. For this reason, and due to the low sample size, a non-parametric test was used to compare the outer loop validation scores between detection methods. Using a Friedman test, there was found to be a significant difference between the performance of the nine different detection methods evaluated $\chi^2(8)=53.2$, $p<0.001$.

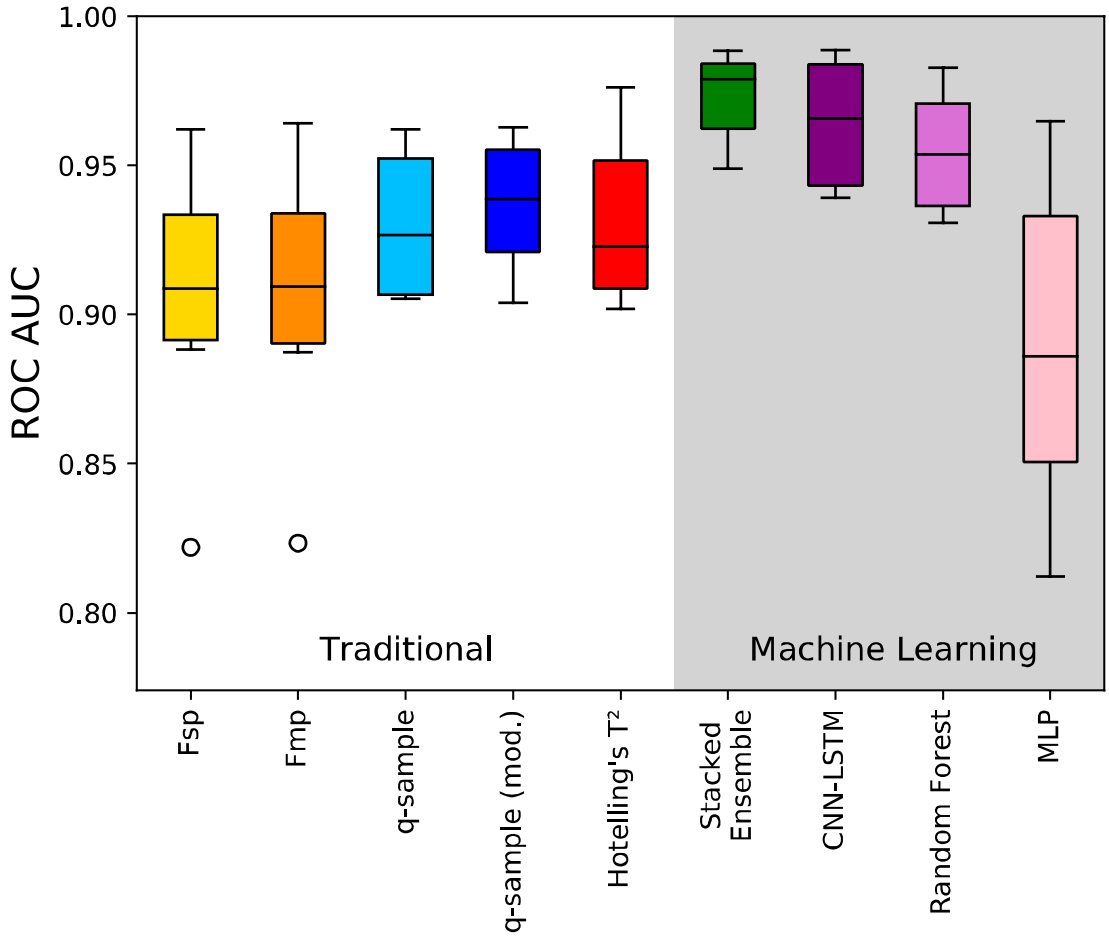

Figure 4-9 Training set cross-validation scores. The ROC AUC scores were compared across the nine detection methods evaluated. Adapted with permission from Wolters Kluwer Health, Inc.: McKearney RM, Bell SL, Chesnaye MA, and Simpson DM. (2022) 'Auditory Brainstem Response Detection Using Machine Learning: A Comparison With Statistical Detection Methods', *Ear & Hearing*, 43(3), pp. 949–960, doi: 10.1097/AUD.0000000000001151.

Post hoc analysis was performed using the Wilcoxon signed-rank test, making multiple pairwise comparisons between the ABR detection methods evaluated. Correction for multiple comparisons was made using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). The Benjamini-Hochberg method controls the false discovery rate (set to 0.05 in the current study) and is considered to be more powerful compared to methods which control the familywise error rate such as the Bonferroni correction method.

Of the statistical ABR detection methods, the original and modified versions of the q-sample uniform scores test and Hotelling's T^2 test were found to be best, with no significant difference in

performance detected between the three methods. The modified q-sample uniform scores test was found to perform statistically significantly better than both the Fsp and the Fmp tests.

Of the machine learning methods, the stacked ensemble and CNN-LSTM performed the best, with the stacked ensemble achieving the highest mean and median outer loop validation fold ROC AUC score. The stacked ensemble performed statistically significantly better than the random forest and the multilayer perceptron, although not significantly better than the CNN-LSTM. The stacked ensemble and random forest performed statistically significantly better than all of the statistical ABR detection methods evaluated. As the best machine learning method, as determined by nested cross-validation using the training set data (highest mean ROC AUC), the stacked ensemble was selected as the machine learning method to first have its detection criterion established using the threshold set data, before being evaluated on the unseen test set data. The optimised hyperparameters, as identified via a random search, were the same across six of the nine cross-validation loops, suggesting a relatively stable algorithm. This same hyperparameter combination was also identified when applying the cross-validation procedure to the entire training set data in order to identify the optimal hyperparameter combination. This hyperparameter combination was used to train the stacked ensemble on the entire training set data before evaluating the algorithm on the threshold and test sets.

4.3.3 Test Set Specificity Evaluation

An effective ABR algorithm should have a low false positive rate, to ensure that clinicians do not mistakenly report an EEG recording as containing a response when it in fact does not. The target specificity level in the current study was 99%, i.e. a false positive rate of 0.01. Each of the ABR detection methods were evaluated on the 15,000 'response absent' data within the test set. The specificities obtained are shown in Figure 4.3.3 for each of the ensemble sizes evaluated. The stacked ensemble (bootstrapped) and Hotelling's T^2 test were able to consistently (across ≥8/10 ensemble sizes) achieve a specificity within the 95% CI for the expected level. When the stacked ensemble had its critical values set using the separate threshold set of 'response absent' data (rather than using the bootstrap), the specificities were below the 95% CI for 7/10 ensemble sizes. The modified q-sample uniform scores test also had its critical thresholds set using the threshold set of 'response absent' data, and again, the specificities achieved were significantly below the expected range for the larger ensemble sizes (≥500 epochs). The Fsp and the Fmp test specificities were outside of the upper range of the 95% CI across all ensemble sizes. The specificities achieved using the original q-sample uniform scores test (utilising zero-padding) were all significantly below the expected range.

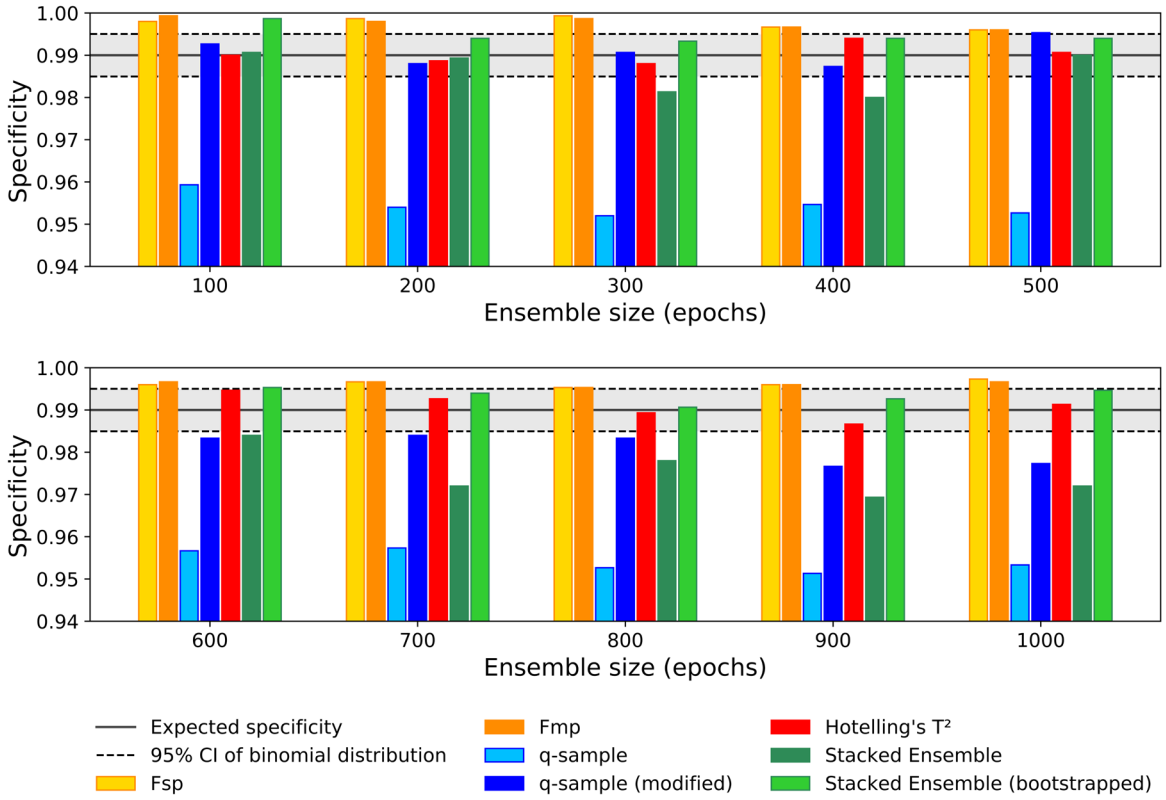

Figure 4-10 Test set specificity evaluation as a function of ensemble size. The specificity of each ABR detection method was evaluated using the 'response absent' data contained within the test set. The expected specificity level and its 95% confidence interval, as calculated from the binomial distribution (n=1,500 trials per ensemble size) are shown. Adapted with permission from Wolters Kluwer Health, Inc.: McKearney RM, Bell SL, Chesnaye MA, and Simpson DM. (2022) 'Auditory Brainstem Response Detection Using Machine Learning: A Comparison With Statistical Detection Methods', *Ear & Hearing*, 43(3), pp. 949–960, doi: 10.1097/AUD.0000000000001151.

4.3.4 Test Set Sensitivity Evaluation

The next step of evaluating the ABR detection algorithms was to assess their sensitivity. In order to do this in a fair manner, the critical values for all of the detection algorithms were adjusted to the level at which they obtained the target specificity level of 99% on the test set 'response absence' data (Chesnaye *et al.*, 2018). This ensured that detection methods with a high false positive rate were not afforded an unfair advantage in terms of sensitivity (Chesnaye *et al.*, 2018). Cochran's Q test found that there was a statistically significant difference in detection rate between the ABR detection methods evaluated, for each ensemble size evaluated, $p<0.001$ (Figure 4-11). Post hoc comparison for detection performance at each ensemble size was performed using the pairwise McNemar test, with a correction for multiple comparisons applied using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). The stacked ensemble

(both the bootstrapped version and the version whose detection criterion was set using the threshold set data) performed statistically significantly better than all of the other detection methods, across all of the ensemble sizes evaluated, on the previously unseen 'response present' test set data (largest adjusted $p < 3\times10^{-6}$) (McKearney *et al.*, 2022). Of the statistical detection methods evaluated, the modified q-sample uniform scores test and the Hotelling's T^2 test achieved the highest detection rates across ensemble sizes, except from for the ensemble size of 100 epochs, where the Hotelling's T^2 test performed least well. For ensemble sizes of 300 to 1,000 epochs, both versions of the q-sample uniform scores test and the Hotelling's T^2 test were found to perform statistically significantly better than both the Fsp and the Fmp.

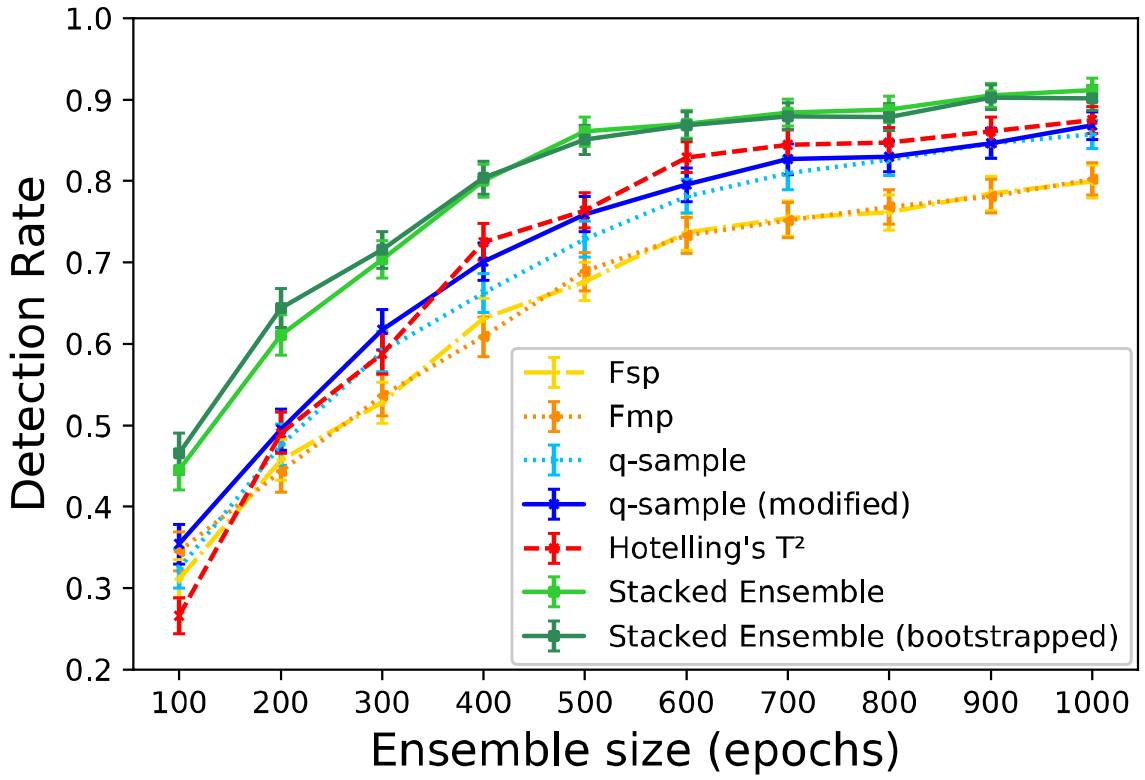

Figure 4-11 Test set sensitivity evaluation. The stacked ensemble (both the bootstrapped version and the version whose detection criterion was set by the threshold set data) had a higher detection rate than all of the other ABR detection methods evaluated. The critical values for each detection method were adjusted to achieve a target false positive rate of 0.01. Error bars represent the 95% CI of the expected binomial distribution centred around each point. Adapted with permission from Wolters Kluwer Health, Inc.: McKearney RM, Bell SL, Chesnaye MA, and Simpson DM. (2022) 'Auditory Brainstem Response Detection Using Machine Learning: A Comparison With Statistical Detection Methods', *Ear & Hearing*, 43(3), pp. 949–960, doi: 10.1097/AUD.0000000000001151.

To evaluate the effects of training set sample size on detection performance, a learning curve analysis was conducted. The results of this are provided in Appendix B.

4.3.5 Analysing ABR Detection Performance by SNR

The test set data were simulated based on the estimated SNRs of the ABR signal within the subject recorded data. In order to provide further in-depth analysis of ABR detection by SNR, to help bring meaning to the detection rates achieved, the test set data were resimulated by taking the 12 ABR templates used to simulate the test set data, rescaling them to obtain the desired range of SNRs, and adding them to the 1,500 test set 'response absent' data (1,000 epochs per ensemble) (Figure 4-12).

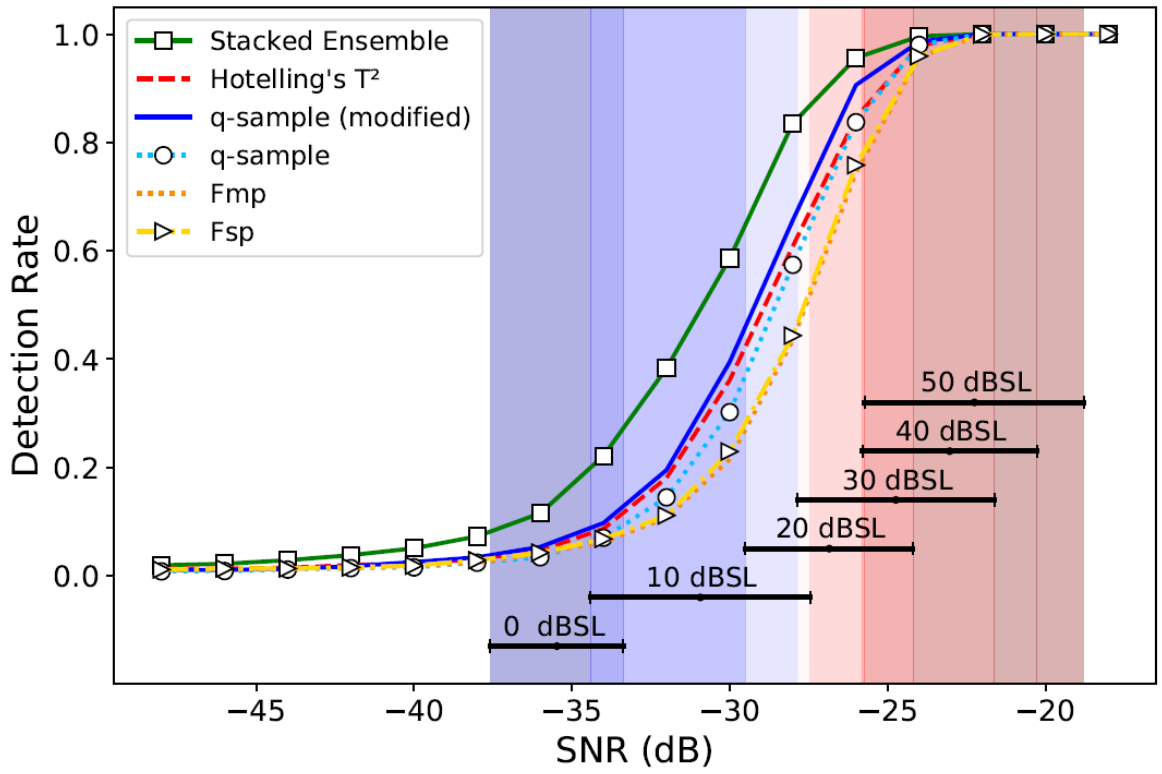

Figure 4-12 The ABR detection rate is shown as a function of SNR. The estimated SNRs (mean ± standard deviation) of the subject recorded data at each stimulus level are superimposed on the figure to provide clinical relevance to the detection performance of the ABR detection methods evaluated. The detection criterion of each detection method was adjusted to the level at which a false positive rate of 0.01 was obtained. Adapted with permission from Wolters Kluwer Health, Inc.: McKearney RM, Bell SL, Chesnaye MA, and Simpson DM. (2022) 'Auditory Brainstem Response Detection Using Machine Learning: A Comparison With Statistical Detection Methods', *Ear & Hearing*, 43(3), pp. 949–960, doi: 10.1097/AUD.0000000000001151.

The stacked ensemble (without bootstrapping) achieved a notably higher detection rate compared to the statistical detection methods evaluated across the range of -35 to -25 dB (Figure 4-12). The stacked ensemble was able to detect responses with the same detection rate as the best-performing statistical detection method (the modified q-sample uniform scores test) when the SNR was almost 2 dB lower, i.e. the sigmoid detection curve of the stacked ensemble was shifted just under 2 dB to the left of the detection curve of the best-performing statistical ABR detection method. Note that the critical value for each detection method was chosen to achieve a set false positive rate of 0.01, ensuring a fair comparison across detection methods (Chesnaye *et al.*, 2018).

4.4 Discussion

The main aims of this study were to identify a machine learning algorithm able to detect the presence or absence of an ABR, and to compare this algorithm with prominent statistical detection methods. Nested k-fold cross-validation using the training set data identified a stacked ensemble as the best machine learning method evaluated. In-depth comparison of the stacked ensemble algorithm with statistical detection methods was performed using the separate test set data. Overall, the results were promising, indicating that machine learning methods can significantly outperform conventional statistical ABR detection methods, using the present dataset. This study also provides methods for training, designing, and evaluating machine learning algorithms for ABR detection.

4.4.1 Specificity Analysis

In clinical practice it is important to have a high level of confidence when deciding that an EEG signal does not contain an ABR (British Society of Audiology, 2019c). It is therefore paramount that ABR detection methods are able to achieve a high and reliable level of specificity. Two main methods were utilised to set the critical value of the stacked ensemble in order to achieve the desired false positive rate. The first was to use a separate set of 'response absent' data (the 'threshold set') and to set the detection criterion at the level at which the desired false positive rate was achieved (for each ensemble size). This method was not successful at controlling the false positive rate for the stacked ensemble or the modified q-sample uniform scores test (Figure 4-10). This is likely due to the data in the threshold set not reflecting the characteristics of the data in the test set, due to the relatively small number of participants which contributed to the data in each set. A second method was therefore employed to control the false positive rate: the bootstrap technique (Section 3.3) (Lv, Simpson and Bell, 2007; Chesnaye *et al.*, 2018). Using the bootstrap technique, the stacked ensemble was able to achieve a specificity level within the 95%

CI of the expected binomial distribution across 8/10 ensemble sizes. Using the bootstrap means that a separate set of data was not required to set the detection criterion as this was determined by the bootstrap on a case-by-case basis for each individual ensemble being evaluated (Chesnaye *et al.*, 2018).

The Fsp and Fmp achieved higher-than-expected specificity levels across all ensemble sizes, with the critical value selected based on the theoretical distribution of the *F* statistic. This is hypothesised to be due to the conservatively applied five degrees of freedom for the numerator of these *F* statistics (Elberling and Don, 1984; Chesnaye *et al.*, 2018). The q-sample uniform scores test achieved a lower-than-expected specificity across all 10 ensemble sizes. This was thought to be due to the use of zero-padding which violated the assumption made by the test that the q sets of spectral bands analysed are independent (Stürzebecher, Cebulla and Wernecke, 1999). Zero-padding was used as it was found to improve the ROC AUC of the test during optimisation of the statistical detection methods. Removal of zero-padding led to the specificity levels falling within, or just above, the 95% CI across ensemble sizes, albeit at the expense of a reduced detection rate when adjusting the detection criterion to a level which achieved a false positive rate of 0.01.

In the case of ABR detection, the term 'false positive' is presently used to refer to the case where an EEG signal is mistakenly determined to contain and ABR. This is highly undesirable. For a newborn hearing screening test, a false positive result would mean an unnecessary referral of a baby for additional specialist testing. Reducing the false positive rate, or at least maintaining a consistent and low false positive rate, avoids unnecessary onward referrals for testing, saving health services administrative and clinical time, as well as avoiding unnecessary stress for parents/carers whilst they wait for an appointment to establish the child's true hearing level. In the case of diagnostic ABR testing, a false positive result could lead to clinicians interpreting the data incorrectly, with the patient's hearing threshold appearing better than it actually is. This error would negatively impact the clinical decision-making process regarding hearing habilitation. The high and stable level of specificity achieved by the bootstrapped stacked ensemble demonstrates that the bootstrap technique can be used to harness the detection ability of machine learning algorithms. There are numerous subject characteristics which may affect the ABR, e.g. age (Hall, 2007). Prematurely born newborns, newborns born at full term, and adults, may exhibit different ABR characteristics, e.g. due to differences in spectral content (Eggermont *et al.*, 1996). This could result in varying false positive rates based on the individual subject characteristics of the patient and the statistical detection method used (Lv, Simpson and Bell, 2007). Using the bootstrap method is expected to result in data from all population groups having an equal chance of a false positive result as the critical value is derived from each separate recording (Lv, Simpson and Bell, 2007).

4.4.2 Sensitivity Analysis

The sensitivity analysis performed (Figure 4-11) showed that the stacked ensemble performed statistically significantly better than the statistical detection methods evaluated, across all ensemble sizes. This good detection performance was also observed in the SNR analysis (Figure 4-12), where the stacked ensemble was able to achieve a better detection rate than statistical detection methods at SNRs that would correspond to behavioural hearing thresholds (0 dB SL). At high SNRs, all ABR detection methods perform well. However, when the SNR is lower, detection naturally becomes more difficult. A better-performing ABR detection method will be better at being able to resolve the ABR threshold.

The results obtained suggest that machine learning algorithms may have the potential to be able to assist clinicians in determining electrophysiological hearing thresholds. Whilst the improvement in detection performance may be considered modest, if applied to say a national newborn hearing screening programme, the effect may be amplified at the population level. The stacked ensemble algorithm, when trained on a larger database of subject recorded clinical data (from both individuals with normal hearing and individuals with a hearing loss) reflecting the intended population for use, may be potentially useful in both evoked potential software to assist clinicians and also in ABR screening devices. Training clinicians on how the algorithm works and how to incorporate its outputs into their clinical decision-making process would be advisable in order to help develop confidence and trust in machine learning algorithms. Further validation using clinical data of the proposed model is first required.

ABR detection performance in previous machine learning studies has been reported using a wide of outcome measures. Typically, most studies report the accuracy achieved (Alpsan, 1991; Acir, Özdamar and Güzeliş, 2006; Davey *et al.*, 2007; Rahbar *et al.*, 2007; Acir, Erkan and Bahtiyar, 2013). Acir, Özdamar and Güzeliş (2006), as in the present study, examined sensitivity and specificity performance, achieving 99.2% sensitivity and 94.0% specificity. The sensitivity performance of the algorithms in the present study was compared using a critical value fixed at the level which achieved a specificity of 99%. This allowed a fair comparison of the different methods at the high level of specificity performance that is required in clinical practice (British Society of Audiology, 2019c). Performance between studies will not only rely on the ability of the proposed detection algorithms, but perhaps to a larger extent on the SNR of the data. Publishing datasets to compare performance between machine learning algorithms, e.g. Kaggle (2021), would make comparison easier. The breakdown of the performance of detection algorithms by SNR (for a fixed false positive rate) makes comparison more feasible. Of the statistical ABR detection methods assessed, the modified q-sample uniform scores test and the Hotelling's T^2

test performed the best (Figure 4-11, Figure 4-12). The reliable specificity performance of the Hotelling's T^2 test, coupled with its good sensitivity and straightforward implementation, makes this method a good benchmark by which future studies may compare new ABR detection algorithm performance.

4.4.3 Limitations and Future Work

Significant efforts were made to simulate the data in a manner which was realistic, however, it is acknowledged that the databases which were used to simulate these data contained data from a limited number of participants. The test set data were derived from ABR recordings from three individuals. The findings of this study must therefore be interpreted with caution, with the current study acting as a proof of concept that machine learning models may be trained to detect the ABR. With increased computational resources, cross-validation could also be applied to the whole train/test procedure as well, randomly allocating different participants to the test set in each iteration. Further research is warranted to further evaluate machine learning algorithms on large amounts of data recorded from individuals with normal hearing and individuals with hearing loss. These individuals should be reflective of the intended target clinical population for which any automated detection algorithm is intended to be used for, e.g. neonates (both with and without a hearing loss). This is necessary in order to validate the proposed algorithm. Learning curves (Appendix B), based on simulated data, suggest that a training set size of more than 900 instances may be expected to achieve a test set score above that of the 95 CI of the best statistical ABR detection method. Simulation and the frequency domain bootstrap may be useful tools to increase the training set size and help improve the generalisable performance of machine learning algorithms.

Whilst efforts were made to optimise the statistical detection methods evaluated, there are always additional parameters and parameter values that may be explored, and so it is possible that the statistical detection method performance could be further enhanced (Chesnaye *et al.*, 2018). For example, the analysis windows used for each detection method could be optimised, as well as further optimisation the frequency components used in the q-sample uniform scores test (Chesnaye, 2019). Additionally, there is a nearly unending list of potential machine learning algorithms, hyperparameter combinations and input features to fine tune and it is possible that another combination of these may perform better than the combination used in this study. It is also quite possible, depending on the dataset used, that the optimal algorithm and hyperparameter combination may be different.

Whilst the bootstrap technique was able to reliably control the false positive rate of the stacked ensemble, it comes at the expense of increased computational cost. For example, for each EEG ensemble being analysed, a single prediction using the stacked ensemble would take ~0.5 seconds using a laptop. However, when applying the bootstrap, an additional 500 predictions were made to estimate the null distribution of the model output (taking around four minutes). This computational cost may certainly be readily reduced by simplifying the model structure, reducing the number of input features to minimise redundancy, streamlining the code (run presently in Python), and reducing the number of bootstrap samples (at the expense of reduced p-value resolution).

4.5 Conclusions

This study showed that a stacked ensemble machine learning algorithm was able to achieve a higher ABR detection rate than prominent statistical ABR detection methods, whilst achieving a reliable specificity level using the bootstrap method. The frequency domain bootstrap and simulation may be used to enhance the size of the dataset to improve model performance and allow sufficient data for model evaluation, whilst allowing the ground truth of the data to be known. Further research is required to evaluate the presented algorithm on a large amount of subject recorded data to assess whether the findings presented in this study will generalise to clinical practice. Successful performance on the target clinical population would need to be demonstrated prior to the methods presented being implemented for clinical use in an evoked potential measurement system.

Chapter 5 Automated ABR Detection and Weighted Averaging

5.1 Introduction

The amplitude of the ABR relative to the background noise in the EEG recording, i.e. the SNR, is very low. This makes detecting the ABR extremely challenging. Background noise in the recording may arise from a number of sources, including background EEG activity, myogenic artefact, ocular movement, and environmental interference, e.g. mains artefact. The primary means by which the SNR is improved is through recording repeated measurements, known as recording epochs, which are typically then combined with equal weighting through coherent averaging to produce a single averaged waveform to be interpreted by the clinician by visual inspection. Being an evoked potential, the ABR is considered to be deterministic, i.e. the evoked potential signal is identical across all recording epochs (Elberling and Don, 1984). Elberling & Don (1984) provide the following equation which describes how the deterministic evoked potential signal (s) and the background noise (v) sum together to form each recorded epoch (x), over each point in time (t):

$$x_{(t)} = s_{(t)} + v_{(t)} \tag{5.1}$$

Evoked potential data are typically arranged as an ensemble matrix of N rows of recording epochs by M sample points of columns (Chesnaye, 2019):

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ x_{N,1} & \cdots & \dots & x_{N,M} \end{bmatrix}$$

The individual recording epochs can be combined into an unweighted average using the following equation (Lyons, 2010):

$$\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i\,(t) \tag{5.2}$$

Assuming that the evoked potential signal is deterministic, the coherent average across the N recording epochs is therefore considered to be:

$$\bar{x}_{(t)} = s_{(t)} + \bar{v}_{(t)} \tag{5.3}$$

with the diacritic referring to the coherent average across the N recording epochs for each point in time (t) (Elberling and Don, 1984). As the background noise is assumed to be random, zero

mean, stationary, i.i.d., and independent from the evoked potential signal, conventional unweighted (mean) averaging will improve the SNR by a factor of N recording epochs (see Section 3.1.1.1) (Wong and Bickford, 1980; Elberling and Don, 1984). This relies on the assumption that the noise is wide-sense stationary, i.e. the probability distribution of the stochastic noise process does not fluctuate over time (Van Drongelen, 2018), and therefore that the mean and variance of the noise is the same across each recording epoch. However, this assumption is often not met, with the variance of the background noise fluctuating over time, e.g. due to changing spontaneous background EEG activity, myogenic activity or changes in the environmental recording conditions meaning that the noise is non-stationary (Hoke *et al.*, 1984; Bataillou *et al.*, 1995). Whilst the use of artefact rejection in recording equipment software may remove recording epochs where the noise exceeds a pre-specified threshold, periods of high noise activity which do not exceed the artefact rejection limit will be incorporated into the coherent average with equal weighting to recording epochs recorded in periods of low noise activity. Weighted averaging, whereby epochs containing less noise (and therefore a higher SNR for 'response present' data) are given greater emphasis relative to those containing more noise, has been proposed to overcome this shortcoming of conventional unweighted averaging (Elberling and Wahlgreen, 1985). As a result, weighted averaging should provide a greater SNR within the weighted average waveform (for 'response present' data) compared to the unweighted coherent average for non-stationary EEG. Weighted averaging may be used to help reduce the recording time required to achieve a desired SNR (Lightfoot and Stevens, 2014). Due to weighted averaging reducing the effect of high noise level epochs within the average, the necessity for artefact rejection becomes diminished, however not obsolete (Lightfoot and Stevens, 2014). Whilst weighted averaging is not a machine learning technique (one of the main focusses of this thesis), data pre-processing of the input features is an important first step in classification. Effective pre-processing combined with optimised detection techniques will likely lead to improved overall detection performance. Weighted averaging reduces the dimensionality of the EEG data from a matrix of N recording epochs by M sample points to a vector of M samples, i.e. feature extraction.

5.1.1 Weighted Averaging

Weighted averaging can be achieved using a variety of implementations and this section shall review some of the various methods presented in the literature. Numerous approaches exist including: weighting individual epochs or blocks of epochs inversely proportional to a measure related to the estimated variance of the noise (Hoke *et al.*, 1984; Elberling and Wahlgreen, 1985; Davila and Mobin, 1992; Riedel, Granzow and Kollmeier, 2001), Kalman weighting (Li, Sokolov and

Kunov, 2002; Cone and Norrix, 2015), adaptive weighted averaging (Bataillou *et al.*, 1995), and sorted averaging (Mühler and Von Specht, 1999; Rahne, von Specht and Mühler, 2008). Weighted averaging has also been successfully applied to the ASSR (Dobie and Wilson, 1994; John, Dimitrijevic and Picton, 2001) and Visual Evoked Potentials (VEPs) (Bezerianos *et al.*, 1995; Bhargav N, Viswanatha and Shailesh M L, 2020). In 1984, Hoke *et al*. proposed a form of weighted averaging applied to the ABR whereby each individual epoch was inversely weighted to the noise within the epoch. In order to simplify the implementation of the algorithm (due to the contemporaneous computational limitations), Hoke *et al*. proposed using the maximum value of all of the samples within the epoch as a proxy estimate of the noise level, which was found to have a high degree of correlation with the standard deviation of the samples. Using simulations of an ABR waveform added to background EEG noise, Hoke *et al*. found that their weighted averaging technique approximated the least-mean square estimate of the true evoked potential signal, achieving a lower root-mean-square error (RMSE) than unweighted averaging, whilst maintained the expected amplitude of the signal.

An alternative approach to weighted averaging was suggested by Elberling and Wahlgreen (1985) whereby the recording epochs were weighted in blocks rather than individually, using the a single point noise estimate calculated as:

$$Var(\mathbf{sp}) = Var\begin{bmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{N,j} \end{bmatrix} \tag{5.4}$$

where $\mathbf{sp}$ is a column vector extracted from the column of sample point j of the ensemble matrix. The equation for the weighted averaging method proposed by Elberling & Wahlgreen (1985) is shown below:

$$\hat{\mathbf{x}} = \left(\frac{\bar{\mathbf{b}}_1}{V_1} + \frac{\bar{\mathbf{b}}_2}{V_2} + \cdots \frac{\bar{\mathbf{b}}_L}{V_L}\right) \cdot \frac{1}{T} \tag{5.5}$$

where $\hat{\mathbf{x}}$ is the weighted average, $\bar{\mathbf{b}}_L$ is the coherent average of the L^{th} block, V_L is the estimated variance of the noise in the L^{th} block, and T is the sum of the inverse of the variances across the L blocks, providing a normalising factor. The normalising factor T constrains the sum of the weights to unity, thereby ensuring that the signal amplitude is unchanged by weighted averaging, i.e. no bias error is introduced. Weighting in blocks of epochs (as opposed to individual epochs) is suggested to provide a more reliable estimate of the variance of the background noise as more data points are available for calculating the estimate (Elberling and Wahlgreen, 1985). Each block of epochs is weighted inversely proportionally to the estimated variance of the background noise within that block, i.e. inverse-variance weighting. The noise level is estimated by calculating the

variance down a single column of samples across all of the recording epochs within the block (in the same manner as the denominator in the F_{sp} equation—Elberling & Don, 1984). The block weights are scaled by a factor inversely proportional to the total the sum of the variances, meaning that the weighted average produced contains an unaugmented signal amplitude (i.e. the signal amplitude should be maintained and not increased or decreased by the weighted averaging procedure). This allows a direct comparison between the weighted average and the coherent average to be made (Elberling and Wahlgreen, 1985). A block size of 250 epochs was recommended by Elberling & Wahlgreen (1985), on the basis that this would include sufficient samples within the single point noise estimate for the noise estimate to be reliable (Elberling and Don, 1984). The size of the block of epochs to be weighted should not be so small such that the estimate of the variance of the background becomes unstable, leading to sub-optimal block weights. Neither should the block size be so large that changes in the background noise level are not resolvable, i.e. that both epochs of high and low noise levels are grouped together in the same block and weighted equally. The optimal block size is therefore a trade-off between these two extremes. Elberling & Wahlgreen (1985) elaborate that it would be worthwhile investigating whether the block size can be optimised further and that a much smaller block size may feasibly be more effective yet. Gerull *et al*. (1996) also identify that using too large a block size is not effective for recordings containing noise fluctuations of short duration.

A later study by Don and Elberling (1994) explored the optimisation of block size for weighted averaging. Using 80 ABR recordings from eight individuals, Don & Elberling (1994) performed weighted averaging using four different block sizes (32, 64, 128, and 256 epochs-per-block—Figure 5-1). Rather than using a single point to estimate the noise, they selected several spaced points in each recording epoch. They found that the residual noise level within the averaged waveform was lowest when using the smaller block sizes. Although the effect of the choice of block size parameter was relatively small (up to a ~1.3% decrease in residual noise relative to the largest block size used: 256). The residual noise continued to reduce as the block size decreased to 32 epochs-per-block and the present author believes it would be useful to investigate if the block size could be reduced even further in order to observe further noise reduction. One potential limitation of their study is that the outcome variable (reduction in residual noise) was estimated using the single point method whilst comparing weighted averaging block sizes which also estimated the variance of the noise using a similar method. The use of simulated data, as used by Hoke *et al.* (1984), where the residual noise levels are known definitively would help to overcome this limitation. Using simulations, Riedel *et al*. (2001) evaluated weighted averaging using a range block sizes (1,2,4,8,16,32,64,128,256). Riedel *et al*. (2001) investigated the use of weighted averaging, estimating the variance of the noise contained within each block by

calculating the average of the power of each of the individual recording epochs in the block. A block size of 32 produced the lowest mean residual noise level. The optimal block size decreased further from 32 to four epochs-per-block when iterative weighted averaging was applied. In iterative weighted averaging, the estimated ABR signal is subtracted from the block of recording epochs prior to calculating the weights. This is performed iteratively with the aim of each successive iteration providing a better ABR signal estimate and subsequently allowing a more accurate noise estimate to be calculated. However, due to the small number of recordings analysed the error bars are substantially overlapping with similar performance shown for all block sizes evaluated between 1–256 (iterative weighted averaging). Further research regarding the best method to estimate the noise levels within each block as well as the optimum block size is still required and is a topic explored in this chapter.

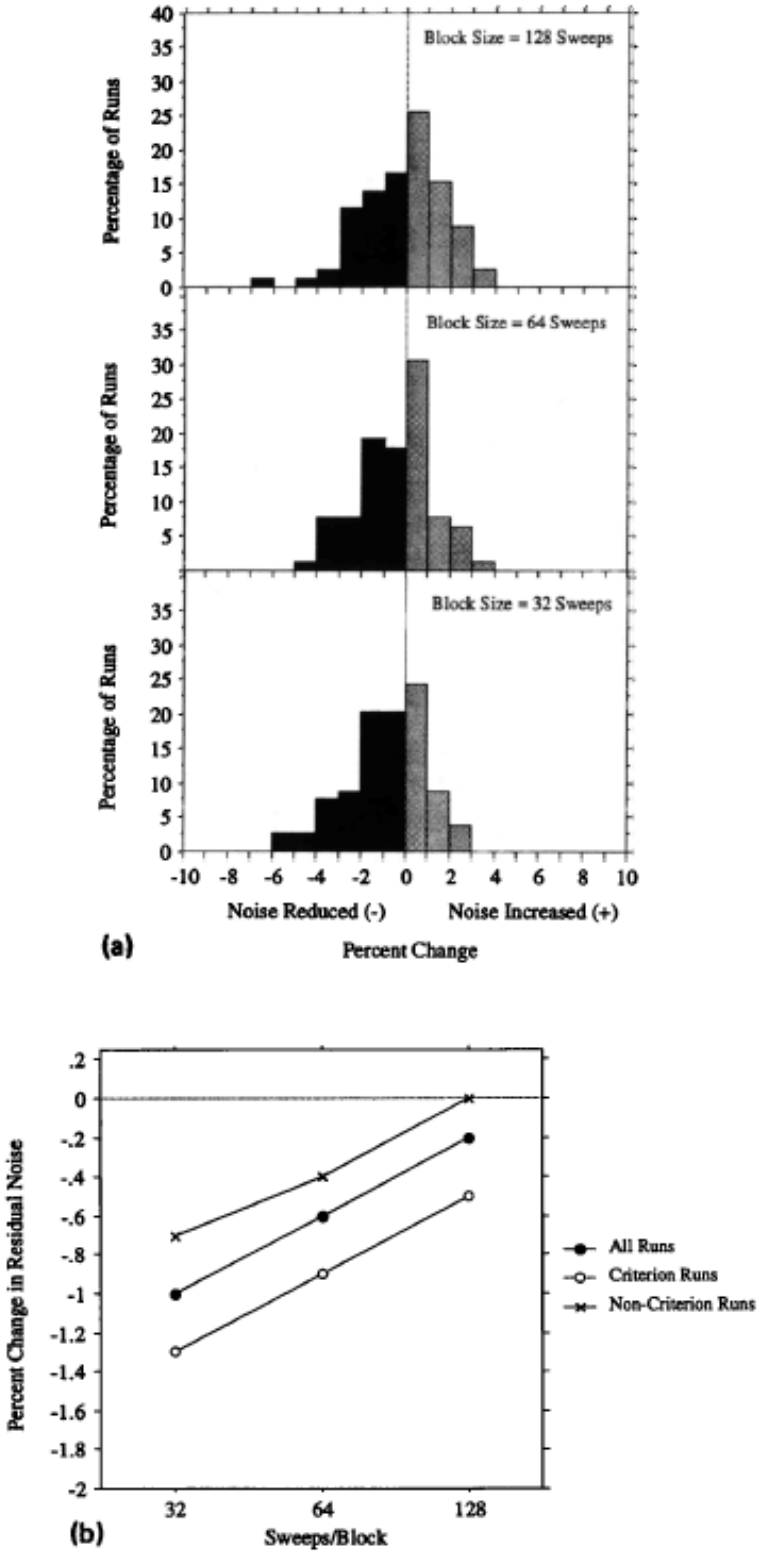

Figure 5-1 In this Figure the term Sweeps/Block is equivalent to the term epochs-per-block used in this study. (a) Histograms showing the percentage change in residual noise using weighted averaging with block sizes of 128, 64, and 32 epochs-per block, relative to using 256 epochs-per-block (the horizontal reference line indicating 0% percentage change in residual noise). (b) The filled circles show the mean percentage change in residual noise across the whole dataset ('All

Runs') using: 32, 64 and 128 epochs-per-block, relative to using 256 epochs-per-block. Performance in Figure (b) is divided into three subsets: 'Criterion Runs' are ABR present ensembles which saw a 5% decrease in residual noise relative to unweighted averaging. 'Non-criterion Runs' saw a <5% reduction in residual noise relative to unweighted averaging. 'All Runs' includes both of these two categories. Reproduced from Don, M. and Elberling, C. (1994) 'Evaluating Residual Background Noise In Human Auditory Brain-Stem Responses', *Journal of the Acoustical Society of America*, 96(5), pp. 2746–2757. doi: 10.1121/1.411281, with the permission of the Acoustical Society of America.

An example of the potential benefits of weighted averaging is shown in Figure 3-1. Here a simulated ensemble of 1,000 epochs was generated by adding an ABR template to a no-stimulus EEG recording. It can be seen that the estimated level of the background noise increased appreciably after 600 recording epochs. This coincided with a degradation in the estimated SNR of the EEG data based on the Fmp detection method (Martin *et al.*, 1994). A form of weighted averaging based on the Elberling & Wahlgreen method (1985) was used to weight blocks of 50 epochs in a simulated ABR ensemble of 1,000 epochs. It can be seen how weighted averaging led to a weighted average (green) which more closely reflected the true ABR signal (black) compared to the conventional (unweighted) averaging method (red).

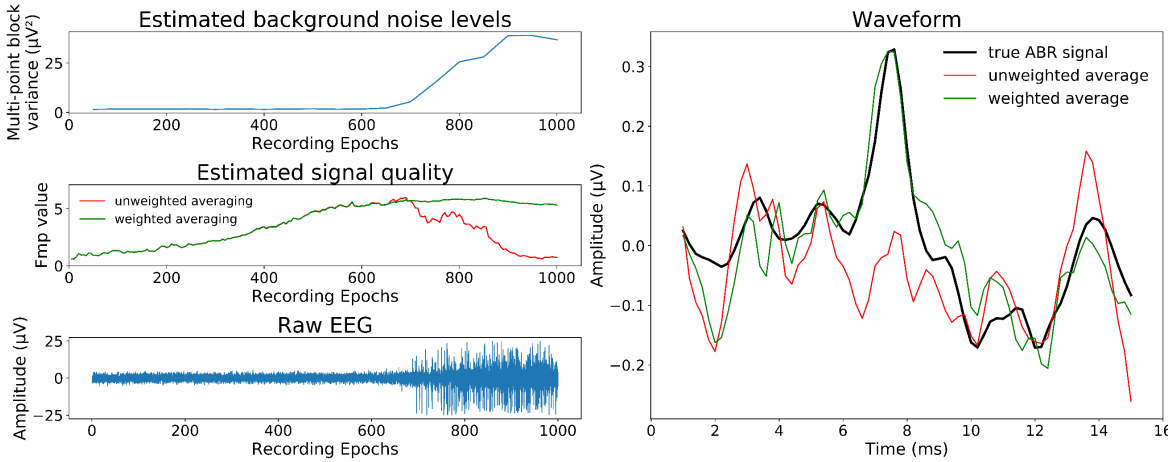

Figure 5-2 Weighted averaging—an example. Note how the estimated signal quality (within the averaged waveform), as estimated by the Fmp, began to decrease after ~700 recording epochs when using unweighted averaging. This was not the case for weighted averaging where the 'noisy' recording epochs were incorporated into the average with lower weight.

5.1.1.1 Alternative Approaches—Kalman Filtering

Another approach to improving the SNR within the ABR waveform is Kalman filtering (Kalman, 1960; Li, Sokolov and Kunov, 2002), which again aims to improve the quality of the estimate of the

signal. Kalman filtering minimises the probability of error in the estimated signal, providing greater confidence in the estimate, and thus improve the SNR (Hall, 2007; Cone and Norrix, 2015). Kalman filtering is recursive, i.e. the information from the previous state (the estimated voltage of the evoked potential in this case) is used in conjunction with ongoing measurements to update the prediction in an iterative two-step process (Figure 5-3). Step one of the Kalman filter process is to 'make an *a priori* estimate of the state of the system' (Van Drongelen, 2018). In the second step, a new *a posteriori* estimate is computed, by fusing the information from the *a priori* estimate and the new measurement (Van Drongelen, 2018).

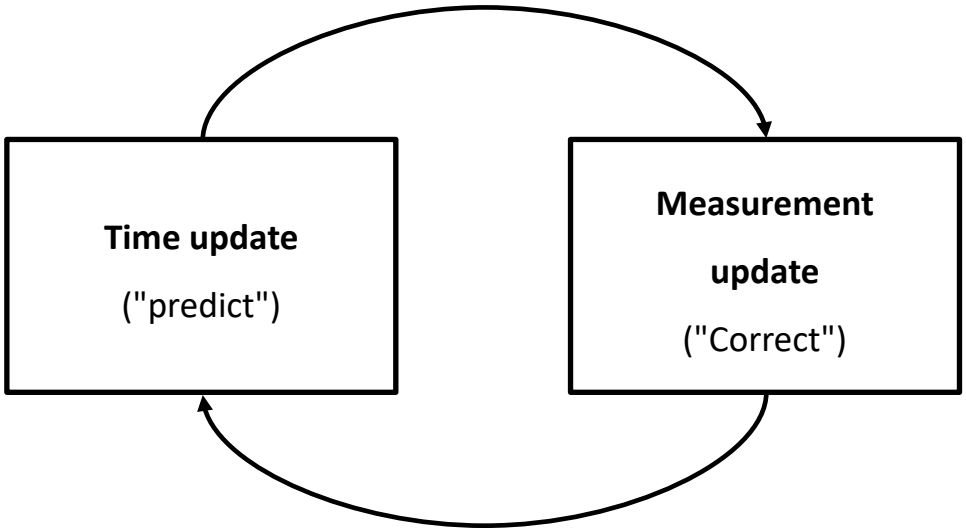

Figure 5-3 The Kalman filter cycle. Figure redrawn, with changes made, based on a figure in Rao, G. M. Nandyala, S. P. and Satyanarayana, C. (2014) 'Fast Visual Object Tracking Using Modified kalman and Particle Filtering Algorithms in the Presence of Occlusions', *International Journal of Image, Graphics and Signal Processing*. MECS Publisher, 6(10), pp. 43–54. Available at: https://doi.org/10.5815/IJIGSP.2014.10.06, under the terms of the CC BY 4.0 license

At each time point, the projected estimate of the current state of the system is combined with the measurement in a weighted average (Khodarahmi and Maihami, 2022). As the ABR signal is considered to be deterministic, Kalman filtering, based on MMSE (Van Drongelen, 2018), will likely produce very similar results to weighted averaging, which also minimises the mean squared error (Elberling and Wahlgreen, 1985). The advantages of Kalman filtering are likely to be more evident in applications involving signals which vary dynamically over time (Kalman, 1960); however, this is not the case for the assumedly deterministic ABR signal. Kalman filtering may potentially be useful for surgical monitoring applications where the goal is to detect a change in the evoked potential signal over time (Hu *et al.*, 2015), in order to help preserve physiological function. One difference between the two methods is that with Kalman filtering being recursive,

there is less data to store in computer memory (Caceres, Sottile and Spirito, 2009). A limitation of Kalman filtering as applied to ABR data is that it is not readily amenable to the application of the most prevalent ABR statistical detection methods (F_{sp}, F_{mp}, Hotelling's T^2 test, and the q-sample uniform scores test). Other forms of weighted averaging produce both a weighted ensemble and a final weighted average, allowing the application of these tests, whereas Kalman filtering does not. The result of Kalman filtering is a final signal estimate rather than a weighted ensemble.

5.1.2 Weighted Averaging—Technical Considerations

Some methods calculate the averaging weights (w) based on a statistic such as the variance across the discrete-time samples within an epoch:

$$w_i = \frac{1}{Var(\mathbf{x}_i)} \tag{5.6}$$

where $\mathbf{x}_i$ is the i^{th} recording epoch in the ensemble. This has the limitation of also including information regarding the evoked potential signal energy as well as the noise energy which is the primary focus. As the SNR of the ABR within the continuous EEG is very low, this has not been considered to have a significant impact (Sörnmo and Laguna, 2005). The ABR signal is assumed to be independent of the background noise (Elberling and Don, 1984):

$$Var(\mathbf{x}_i) = Var(\mathbf{s}_i) + Var(\mathbf{v}_i) \tag{5.7}$$

where $\mathbf{s}$ is the evoked potential signal vector and $\mathbf{v}$ is the background noise vector in recording epoch $\mathbf{x}$. Lütkenhöner *et al*. (1985) and Fan & Wang (1992) observed that the use of weighting strategies based on the variance of the samples within an epoch can lead to an underestimation of the amplitude of the signal. This has been hypothesised to be a result of the weights (not normalised) being influenced by the evoked potential signal energy within the noise estimate (Fan and Wang, 1992; Gerull, Graffunder and Wernicke, 1996). If the weights are not constrained by the sum of the variables used as the noise level estimates, the signal may be distorted and no longer comparable to the coherently averaged waveform. This will negatively impact visual interpretation by clinicians who make judgements regarding the evoked potential signal amplitude to inform their interpretation (British Society of Audiology, 2019c). Lütkenhöner *et al*. (1985) and Gerull *et al*. (1996) proposed alternative methods of avoiding signal waveform underestimation by increasing the degrees of freedom of the noise variance estimates. These include widening the analysis window and therefore the number of samples from which the noise variance is estimated (Lütkenhöner, Hoke and Pantev, 1985), and the use of filtering to pre-whiten the EEG (Lütkenhöner, Hoke and Pantev, 1985; Gerull, Graffunder and Wernicke, 1996). By increasing the degrees of freedom of the noise variance estimate, the accuracy of the weighting

factor can be improved as a greater number of independent datapoints are available, thus avoiding misestimation of the signal.

5.1.2.1 Weight Normalisation

Elberling & Wahlgreen (1985) suggested that each weight incorporate the inverse of the total sum of the inverse of the variances calculated, thereby normalising each weight (Equation 5.5). On the basis that the evoked potential signal is deterministic across all epochs, and therefore that the signal is identical in each epoch, all of the noise estimates are scaled by this normalising factor to sum to unity. No misestimation of signal amplitude should therefore occur as the amplitude scale has been corrected (Gerull, Graffunder and Wernicke, 1996; Kumaragamage, Lithgow and Moussavi, 2016). This allows the weighted average to be compared to the unweighted average (Elberling and Wahlgreen, 1985).

5.1.2.2 The Effects of Weighted Averaging on Statistical ABR Detection Methods

Signal processing techniques including weighted averaging are known to have effects on the statistical properties of the EEG data (Hoke *et al.*, 1984; Elberling and Wahlgreen, 1985; Lütkenhöner, Hoke and Pantev, 1985; Gerull, Graffunder and Wernicke, 1996), with an increase in residual high-frequency components (relative to conventional averaging) due to the weighting factor being predominantly influenced by low-frequency components (Hoke *et al.*, 1984; Elberling and Wahlgreen, 1985). Due to the sometimes unpredictable effects of weighted averaging, Lütkenhöner *et al*. (1985) warned that 'blind confidence in the method is dangerous'. The effects of any alterations to the statistical properties of the data on the statistical ABR detection methods used clinically have not been previously evaluated. Given the use of weighted averaging in clinical evoked potential devices, this topic merits further consideration including quantification of the problem as well as consideration of methods to mitigate any effects on detection method performance.

5.1.3 Formulation of the Research Problem

Weighted averaging has been shown to be an effective technique at improving the SNR of ABR recordings (Hoke *et al.*, 1984; Elberling and Wahlgreen, 1985; Gerull, Graffunder and Wernicke, 1996; Riedel, Granzow and Kollmeier, 2001; Cone and Norrix, 2015). One of the more prominent methods is the form of weighted averaging proposed by Elberling & Wahlgreen (1985) whereby groups of epochs are weighted together as a block in order to provide a more accurate estimation of the noise. Don & Elberling (1994) went on to show that using smaller block sizes, down to a size of 32 epochs-per-block, helped to reduce the residual noise level in the coherent average. Riedel,

Granzow and Kollmeier (2001) also found a block size of 32 to reduce the residual noise level most effectively in the averaged waveform (4 epochs-per-block when using iterative weighted averaging). It is possible that the observed benefit would continue to increase using even smaller block sizes and the optimal block size is not known. This parameter likely has a prominent impact on the effectiveness of a weighted averaging algorithm. Evidence-based recommendations for a value for the block size parameter are scarce in the literature and may improve ABR detection performance further if the SNR within the coherent average can be enhanced. It is acknowledged that these theoretical improvements may be marginal, however, at the population level small improvements, e.g. in the performance of a nationwide screening test, may have a meaningful impact. An additional feature requiring further research is the method by which the noise levels are estimated within each block (Section 5.1.2). Obtaining a better estimate of this would improve the accuracy of the weights and therefore likely increase the effectiveness of weighted averaging by allowing smaller block sizes to be used.

Data processing techniques such as weighted averaging have the potential to alter the statistical properties of the data (Lütkenhöner, Hoke and Pantev, 1985). It is not known how these affect the performance of the detection methods commonly used in clinical evoked potential devices (the Fsp/Fmp). Weighted averaging will affect not only the signal estimate of the Fmp equation, calculated from the weighted average, but also the background noise estimate from the weighted ensemble. Research into the effects of weighted averaging on ABR test performance has typically focussed on the reduction of residual noise, the increase of the SNR for 'response present data, or a detection measure statistic in 'response present' data (Elberling and Wahlgreen, 1985). The results in the literature are often anecdotal, using a handful of EEG recordings. Research using a large amount of data in order to observe the effects of weighted averaging on the sensitivity and specificity of detection methods has not been performed. This work is required in order to inform the use of statistical detection methods in conjunction with weighted averaging.

5.1.4 Aims and Objectives

Aim 1. To optimise weighted averaging by identifying the value of the epochs-per-block parameter that reduces noise within the averaged waveform and improves ABR detection the most.

Objective 1a: Compare a range of epoch-per-block parameter values across the range 1–1,000 (equivalent to unweighted averaging in the case of a 1,000-epoch ensemble).

Objective 1b: This parameter will be specifically evaluated across a range of performance measures including reduction in residual noise level, effects on Fmp value, and ABR detection.

Aim 2. Compare methods of estimating the variance of the noise level within each block, to further optimise weighted averaging.

Objective 2a: Compare the 'multiple points' method of estimating the noise within a block with the 'variance of the whole block' method, specifically focussing on residual noise reduction within the averaged waveform and ABR detection (ROC AUC).

Aim 3. Investigate the effects of weighted averaging on the Fmp statistical ABR detection method.

Objective 3a: For each block size parameter evaluated, compare the effects of weighted averaging on the Fmp value for both 'response present' and 'response absent' data. Evaluate the effects of weighted averaging on sensitivity and specificity.

5.1.5 Chapter-Specific Acknowledgements

Please note that most of the findings of the study presented in this chapter have been published as a journal article:

McKearney, R. M., Bell, S. L., Chesnaye, M. A., and Simpson, D. M. (2023) 'Optimising Weighted Averaging for Auditory Brainstem Response Detection'. *Biomedical Signal Processing and Control*, 83, p. 104676. Available at: https://doi.org/10.1016/j.bspc.2023.104676.

 Regarding author contributions, the manuscript was based on the work conducted as part of this PhD thesis and was written by the present author with feedback provided by all other authors (Prof. Steven Bell, Dr Michael Chesnaye, and Prof. David Simpson). This thesis chapter has been updated to reflect feedback from the co-authors provided on the journal manuscript. Dr Michael Chesnaye helped by discussing ideas regarding the bias observed in the Fmp statistic and its interaction with the weighted averaging procedure. Dr Michael Chesnaye also provided feedback on the journal manuscript, including on the format of some equations in the manuscript to better reflect the number of degrees of freedom present in the noise variance estimates. Prof. Steven Bell and Prof. David Simpson (PhD supervisors of the present author) provided supervisory guidance regarding all aspects of the study.

5.2 Methods

5.2.1 Data

The data used in this study were the same as those described in Section 4.2.1. In summary, these comprised of a no-stimulus EEG database (~6.5 hours of recordings from 15 individuals) (Madsen, 2010; Madsen *et al.*, 2018), and an ABR database containing threshold series recordings from 12 individuals (Chesnaye *et al.*, 2018). These data were band-pass filtered from 30 to 1,500 Hz using a 3rd-order Butterworth filter. These filter settings were chosen to reflect those recommended by the British Society of Audiology (2019c).

5.2.2 Ethics

Overarching ethical approval was granted by the University of Southampton Faculty Ethics Committee to use the data from the subject recorded ABR and no-stimulus EEG datasets for the purpose of secondary data analysis for a range of research activities throughout these PhD studies (ERGO 55576).

5.2.2.1 No Stimulus Data

A set of 'no response' data were constructed by re-constructing the no-stimulus EEG database into a series of continuous recording epochs. These epochs were grouped together into sets of 1,000 recording epochs to form ensembles. The recording epoch duration reflected a simulated stimulus rate of 30.3 Hz to match that used to record the 'response present' data, although naturally no auditory stimulus was used to collect the no stimulus data. A rejection level of ±25 µV was applied. A higher rejection level was used than might be applied clinically as including more noise will likely make it easier to observe any changes in the residual noise level as a result of weighted averaging in order to measure its effects. Artefact rejection levels used in clinical testing are recommended to be up to ±10 µV (British Society of Audiology, 2019c). Using the above procedure, 2,301 ensembles of 1,000 epochs were generated.

5.2.2.2 ABR 'Response Present' Data

The ABR 'response present' data were simulated by making a copy of the 2,301 no stimulus ensembles and adding a single scaled ABR template (from one adult participant with normal hearing) to every recording epoch within all of the 2,301 ensembles. By simulating the 'response present' data it was possible to know definitively the noise levels within each recording epoch as well as the true SNR. This allowed the exact effects of the averaging method used to be observed. The ABR template was given a peak-to-peak amplitude (wave V to SN10) of 500 nV which is at the

higher end of those physiologically observed (Hall, 2007). This amplitude was chosen as it was found to produce a wide spread of SNRs given the high noise levels in the dataset, helping to avoid floor and ceiling effects when evaluating detection. Another reason for this larger choice of ABR signal peak-to-peak amplitude is that one of the methods for estimating the noise levels in each block (the 'VAR Whole Block' method described in Section 5.2.6) is adversely affected by the presence of a large response. Some degree of certainty is required that this method will hold up even when a large response is present.

5.2.3 Analysis Window

All of the calculations performed in this study, e.g. residual noise, SNR, Fmp level and estimated noise level were applied to a fixed analysis window containing the discrete-time samples within 1–15 ms of each recording epoch. This avoids stimulus artefact which may occur shortly after the onset of the stimulus and covers the main latency period expected of the ABR (Chesnaye *et al.*, 2018). This analysis window length corresponds to 71 digitised sample points per recording epoch at the sampling rate of 5 kHz.

5.2.4 ABR Detection Method

As well as evaluating the effects of weighted averaging on the noise levels and the SNR, it is also important to understand the effects of this alteration to the properties of the data on ABR detection methods. Whilst this has been done in previous studies using three ABR recordings as an illustrative example (Elberling and Wahlgreen, 1985), the effect of weighted averaging on ABR detection methods has not been evaluated in depth. For the purposes of this study the Fmp was used (Martin *et al.*, 1994):

$$Fmp = \frac{Var(\bar{\mathbf{x}})}{\frac{1}{N}\left(\frac{1}{Q}\sum_{i=1}^{Q} Var(\mathbf{sp}_i)\right)} \tag{5.8}$$

where Q is the number of chosen single point columns ($\mathbf{sp}$) down the ensemble matrix (i.e. across recording epochs) that are used to estimate the noise level. The number of *chosen* single point columns used (Q) may be multiple or all columns available, i.e. $1 < Q \leq M$. If $Q = 1$, then the *F* statistic calculated would be equivalent to the Fsp as provided by Equation 3.17. This detection method was chosen, out of a number of alternative methods (Chesnaye *et al.*, 2018), as it is implemented in commercially available auditory evoked potential software, making it clinically relevant. The Fmp is also mathematically closely related to its predecessor, the Fsp. Therefore, conclusions drawn when using the Fmp can theoretically be extrapolated to the Fsp.

For the purposes of calculating the Fmp on weighted data: the numerator of the Fmp equation was calculated as the variance of the weighted average (as opposed to the unweighted coherent average), and the denominator was calculated as the mean of the variance taken down multiple points of the weighted recording epochs in the ensemble. The weighted ensemble can be calculated by reorganising Equation 5.5, provided by Elberling and Wahlgreen (1985), and applying it to the raw blocks of weighted EEG data as opposed to the block sub-averages:

$$\breve{\mathbf{X}} = \begin{bmatrix} \mathbf{B}_1 \cdot \left(\frac{L}{V_1 \cdot T}\right) \\ \mathbf{B}_2 \cdot \left(\frac{L}{V_2 \cdot T}\right) \\ \vdots \\ \mathbf{B}_L \cdot \left(\frac{L}{V_L \cdot T}\right) \end{bmatrix} \tag{5.9}$$

where L is the number of separate blocks of recording epochs, $\breve{\mathbf{X}}$ is the weighted ensemble, $\mathbf{B}_L$ is the L^{th} block, V_L is the estimated variance of the noise in the L^{th} block, and T is the sum of the inverse of the variances across the L blocks (a normalising factor). The coherent average of the weighted ensemble $\breve{\mathbf{X}}$ is equal to the weighted average $\hat{\mathbf{x}}$, provided in Equation 5.5, allowing the Fmp statistic to be calculated.

The specificity level of the detection method was evaluated relative to a pre-determined desired false positive rate of 0.01. This high level of specificity was selected as ABR interpretation in clinical practice typically requires a high degree of specificity (i.e. a low false positive rate), due to the negative impact of falsely detected responses on clinical outcomes (British Society of Audiology, 2019c).

5.2.5 Weighted Averaging: Block Size

Weighted averaging applies weights to blocks of epochs based on an estimation of the noise level within each block. One of the key parameters to be evaluated was the number of epochs-per-block. Ideally a block size of one would be used (i.e. epochs being weighted individually), in order to provide weights specific to each recording epoch as noise levels may fluctuate at any point in time (Don and Elberling, 1994). Unfortunately, due to the small number of samples present when using a small block size, the noise level cannot be accurately determined, and the estimated variance of the noise in the block may be inaccurate, leading to sub-optimal weights being calculated. A trade-off is therefore sought between having sufficient epochs within each block to be able to accurately estimate the noise level and applying the weights precisely over time to small numbers of epochs.

In this study a range of values of the epochs-per-block parameter were explored. Given the ensemble size of 1,000 recording epochs, the values explored were all of the factors of 1,000:

$$1,\ 2,\ 4,\ 5,\ 8,\ 10\ ,\ 20,\ 25,\ 40,\ 50,\ 100,\ 125,\ 200,\ 250,\ 500,\ 1000$$

5.2.6 Weighted Averaging: Estimation of the Noise Level

For any given block size, the data within each block is used to estimate the noise level in order to calculate the weights to be applied. Previous studies have estimated the noise level by calculating the variance down a single point of the ensemble (using the same sample point in time across recording epochs) (Elberling and Wahlgreen, 1985). Don and Elberling (1994) recommended calculating the variance using all samples across eight evenly spaced columns in the ensemble. In this study we have combined these methods, using a multiple points method (Martin *et al.*, 1994) which takes the mean of the estimated variance of the noise across multiple time points, i.e. the denominator in the Fmp detection method equation (Martin *et al.*, 1994). Specifically, the variance was computed separately down all 71 columns in the block of recording epochs and the noise level estimate was calculated as the mean of these separate point estimates:

$$V_L = \frac{1}{Q}\sum_{i=1}^{Q} Var\left(\mathbf{sp}_{\boldsymbol{i}}(\boldsymbol{B_L})\right) \tag{5.10}$$

where V_L is the estimated noise level of the L^{th} block of recording epochs, Q is the chosen number of single point columns used, and $\mathbf{sp}_{\boldsymbol{i}}(\boldsymbol{B_L})$ is the i^{th} chosen single point noise estimate from the L^{th} block. By estimating the noise level down columns of the ensemble and then averaging them together, the noise level estimate is theoretically unaffected by the presence of an ABR signal as the response is assumed to be deterministic with identical amplitude across all of the recording epochs within each column. This method is subsequently referred to as the 'VAR MP' method and serves to act as a baseline by which to compare the performance of other noise level estimation methods.

An alternative method of estimating the noise level within the block ('VAR Whole Block') was also considered. In this second method the variance was calculated across all of the concatenated points across all of the recording epochs in the block (an extension of the method proposed by Don and Elberling (1994) where eight columns of points were used):

$$V_L = Var(\mathbf{B}_L) \tag{5.11}$$

where V_L is the estimated noise level of the L^{th} block of recording epochs ($\mathbf{B}_\mathrm{L}$): a matrix with the dimensions $\frac{N}{L}$ by M (the number of recording epochs per block by the number of sample points

per recording epoch). For example, using a weighted block size of two epochs (71 samples in each), the variance of all 142 concatenated discrete-time sample points would be used to estimate the noise level within the recording epochs in that block. This method has the advantage of making use of more data samples in the estimation of the variance of the noise. The limitation, however, is that the noise level estimate is confounded by the presence of the ABR signal. It is anticipated that because the variance of the ABR signal within a recording is low relative to the variance of the noise, that this effect will have a limited impact (Sörnmo and Laguna, 2005).

The equation used for applying the weights to the averaging process was that provided by Elberling & Wahlgreen (1985), shown previously in Equation 5.5.

5.2.7 Additional Simulation of Stationary Data

This study analysed subject recorded no-stimulus EEG data, with 'response present' data being simulated through the addition of an ABR template to the subject recorded no-stimulus EEG data. These data were of varying degrees of stationarity. In order to provide further evaluation on the effects of the presented weighted averaging algorithms on stationary data, a further simulation was performed. For stationary data, unweighted averaging will be optimal, and it is necessary to be aware of any untoward effects weighted averaging may have on these data. For all of the ensembles in the dataset used in the first part of this study, the mean variance of the noise across all 1,000 recording epochs was calculated. The noise in each recording epoch was then scaled so that its variance was made equal to the mean variance of the noise across all recording epochs in the ensemble. If present, the ABR template was readded after scaling the noise.

5.3 Results

5.3.1 Evaluation of Noise Level Estimation Methods

Within each block of recording epochs, the noise level was estimated in order to obtain the weights for weighted averaging. The more accurately the noise level within each block can be estimated, the more effective weighted averaging will be. Two separate methods of estimating the variance of the noise were evaluated in order to calculate the weights for weighted averaging. These were:

1. The mean of the estimated variance of noise calculated separately down each column of the ensemble—the 'VAR MP' method. All 71 sample point columns were used in this calculation.

2. The variance of all sample points in the block concatenated—termed herein as the 'VAR Whole Block' method.

Figure 5-4 shows a comparison of the partial ROC AUC score (Walter, 2005) across a range of block sizes used with weighted averaging. The partial ROC AUC score is the area under a partial region of the ROC curve, in this case the region corresponding to a false positive rate of ≤ 0.05, placing emphasis on the high specificity levels expected in clinical practice. A block size of 1,000 epochs includes all of the epochs in the ensemble and is therefore identical to applying no weighting, i.e. mean coherent averaging. This level may therefore serve as a baseline performance level. As the block size decreases, the performance of both noise estimation methods gradually improved before dropping off at low block sizes. The 'VAR Whole Block' method was able to reach a higher partial ROC AUC, using a lower block size, which may be expected as more sample points are included in the variance estimate when concatenating the entire block, making the noise estimate more reliable. Note that it is not possible to estimate the variance of the noise using the 'VAR MP' method for a block size of one as it is not possible to calculate the sample variance of a single number. The bootstrap method (with 500 bootstrap samples) was used to estimate the distribution of the partial ROC AUC statistic for each block size evaluated, providing standard error values for Figure 5-4. A paired-sample permutation test (Fisher, 1935; Pitman, 1937; Good, 2000) (using 5,000 permutations) was used to evaluate whether there was a difference in detection performance between the 'VAR Whole Block' method and the 'VAR MP' method. Correction for multiple comparisons was performed using the Bonferroni method (Bonferroni, 1936). The 'VAR Whole block' method was found to perform statistically significantly better than the VAR MP method across block sizes of 2-to-10 epochs-per-block.

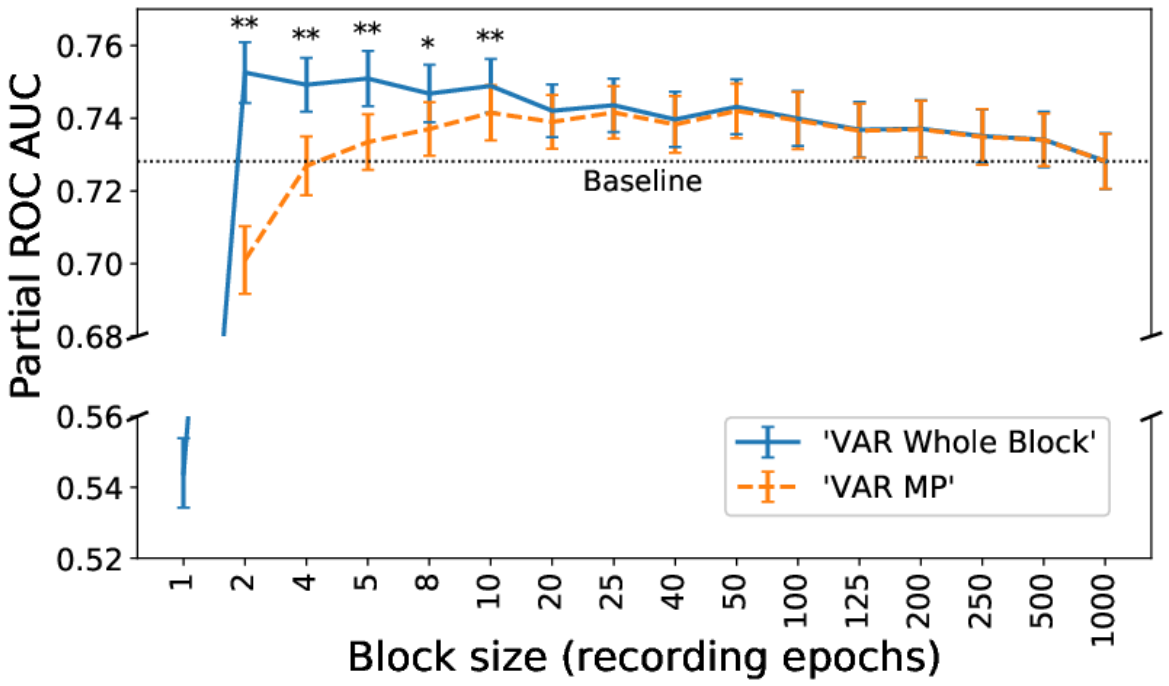

Figure 5-4 Comparison of two methods for estimating the variance of the noise within each block. The evaluation metric used was the partial ROC AUC, i.e. the area under a partial region of the ROC curve, in this case the region corresponding to a false positive rate of ≤ 0.05. A higher partial ROC AUC score corresponds to a better ability to discriminate between 'response present' and 'response absent' data, over the false positive rates of interest. A single asterisk, *, indicates a Bonferroni-corrected two-sided *p* value of < 0.05, as calculated using a paired permutation test. A double asterisk, **, indicates a Bonferroni-corrected two-sided *p* value of < 0.01. Error bars represent the bootstrapped standard error of the partial ROC AUC. Figure reproduced with minor adaptations, in accordance with the CC BY 4.0 license, from McKearney, R. M. et al. (2023) 'Optimising Weighted Averaging for Auditory Brainstem Response Detection', *Biomedical Signal Processing and Control*, 83, p. 104676. Available at: https://doi.org/10.1016/j.bspc.2023.104676.

Figure 5-5 shows the ROC curves for each of the block sizes evaluated using weighted averaging. Figure 5-5 provides information complementary to Figure 5-4, but by presenting the entire ROC curve instead of the ROC AUC, a detailed breakdown of the Fmp performance across a range of false positive rates may be observed. The right graph in Figure 5-5 shows partial ROC curves with a false positive rate of up to 0.05. In this range, the block size of one epoch-per-block was an outlier, performing least well. Aside from this, the general trend was that the smaller the block size used, the better the performance. As block size decreases, weights can be allocated more

precisely to the recording epochs, however, at the expense of a potentially less accurate estimate of the noise level. The performance appears generally to increase until the estimate of the noise level becomes unreliable, e.g. for a block size of one epoch-per-block.

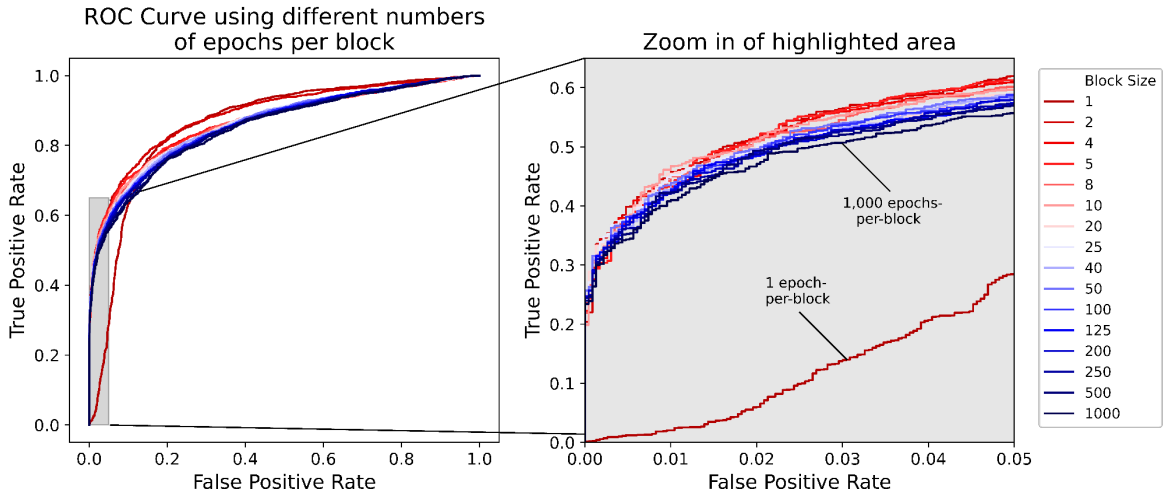

Figure 5-5 Receiver operating characteristic curves for each block size used with weighted averaging. The graph on the right shows the partial ROC curves corresponding to the bottom-left hand corner of the graph on the left. This zoomed in region covers the levels of false positive rate that would typically be required for clinical purposes and is therefore the most relevant region. Whereas Figure 5-4 provides a summary of the area under the curves in the graph on the right, this graph provides a visual breakdown of detection performance by block size across a range of false positive rates, confirming that lower block sizes generally performed better than larger ones across a range of false positive rates.

Figure 5-6 shows that across all block sizes used with weighted averaging, the mean and median residual noise in the averaged waveform using the 'VAR Whole Block' method was always less than or equal to that obtained when using the 'VAR MP' method. Generally speaking, weighted averaging led to lower residual noise levels (calculated as the RMS of the averaged waveform after the ABR signal template, if present, had been removed) within the averaged waveform provided that the block size used was not too low. Note that the mean residual noise levels were greater than the median residual noise levels across all the evaluated block sizes, indicating a positively skewed distribution. This indicates that whilst a block size that produces a lower median residual noise level will be optimal in most cases, there will be a minority of cases where this choice of low block size will be very detrimental to the recording (i.e. increasing the residual noise).

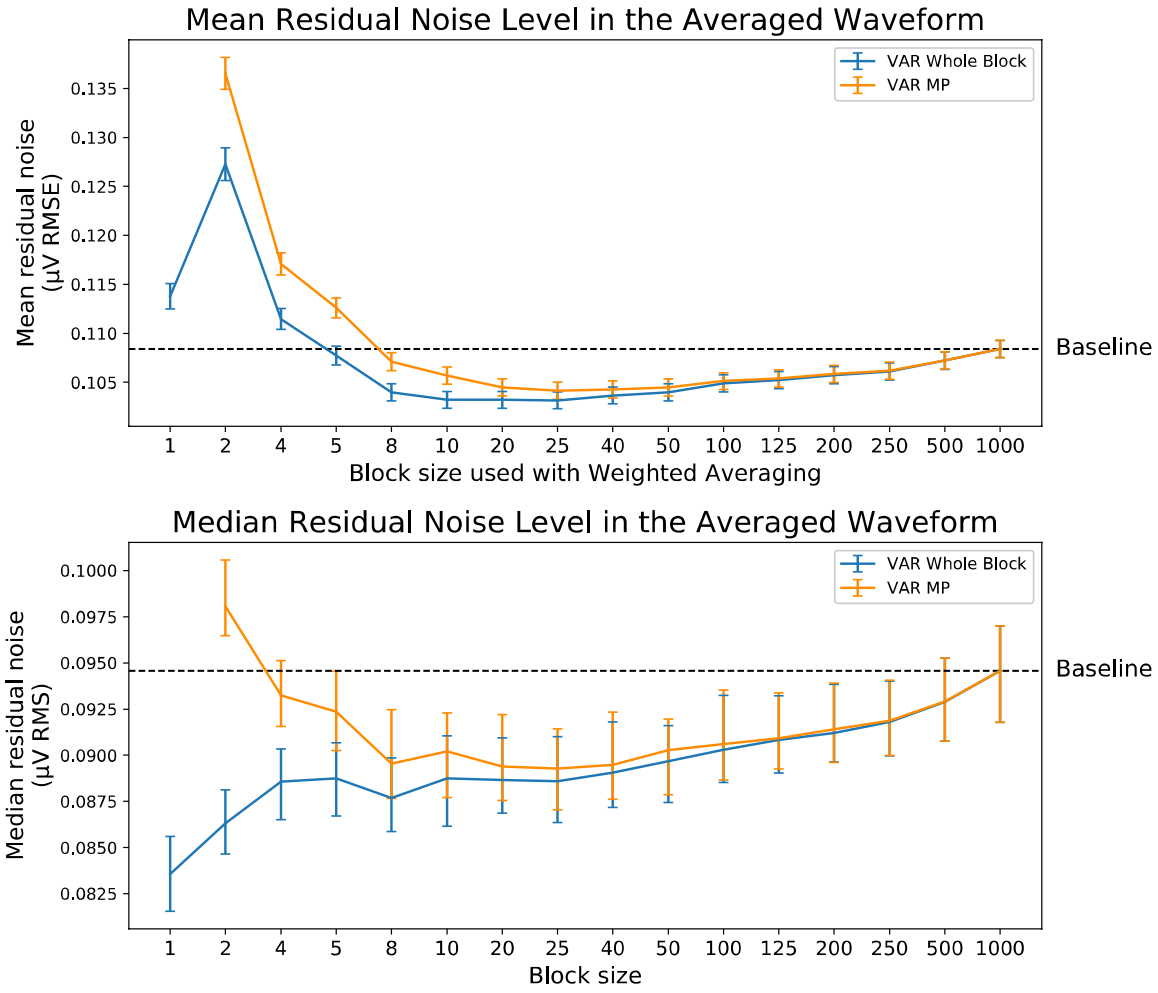

Figure 5-6 Mean and median residual noise levels in the averaged waveform. Figure reproduced with minor adaptations, in accordance with the CC BY 4.0 license, from McKearney, R. M. et al. (2023) 'Optimising Weighted Averaging for Auditory Brainstem Response Detection', *Biomedical Signal Processing and Control*, 83, p. 104676. Available at: https://doi.org/10.1016/j.bspc.2023.104676.

Based on the evidence presented in Figure 5-4 and Figure 5-6, the subsequently presented data were performed using the 'VAR Whole Block' method as the data suggested it to be the better of the two noise estimation methods evaluated. Whilst the 'VAR Whole Block' method potentially makes better use of the available information when estimating the noise level in the block compared to the 'VAR MP' method, it has the limitation of including a bias error (the variance of the ABR signal). This bias error is expected to be small for low SNR signals such as the ABR (Sörnmo and Laguna, 2005). A separate simulation exploring the magnitude of this limitation is shown in Appendix C. In summary: the 'VAR Whole Block' method was able to estimate the variance of the noise at least equally effectively or more so than the 'VAR MP' method across block sizes, provided the SNR was less than approximately -15 dB. Large ABR responses elicited by a 50 dB SL stimulus have SNRs in the range of -34.6 to -22.9 (mean = -27.9, n=12) (Chesnaye,

2019), suggesting that the 'VAR Whole Block' method is favourable for low-SNR evoked potentials such as the ABR.

5.3.2 The Effects of Weighted Averaging on ABR Detection using the Fmp / Evaluation of the Optimal Block Size

5.3.2.1 Fmp Values

This section aimed to investigate the effects of weighted averaging on the Fmp statistic. It is tempting to assume that if residual noise is lower in the weighted average that the Fmp value must therefore be higher, however, this is not necessarily the case as the denominator within the Fmp equation also takes into account the estimated variance of the noise level across all of the weighted recording epochs. It was therefore necessary to explore separately the effects of weighted averaging on the Fmp statistic, including an analysis of the 'response absent' condition where a change in Fmp performance could impact the false positive rate.

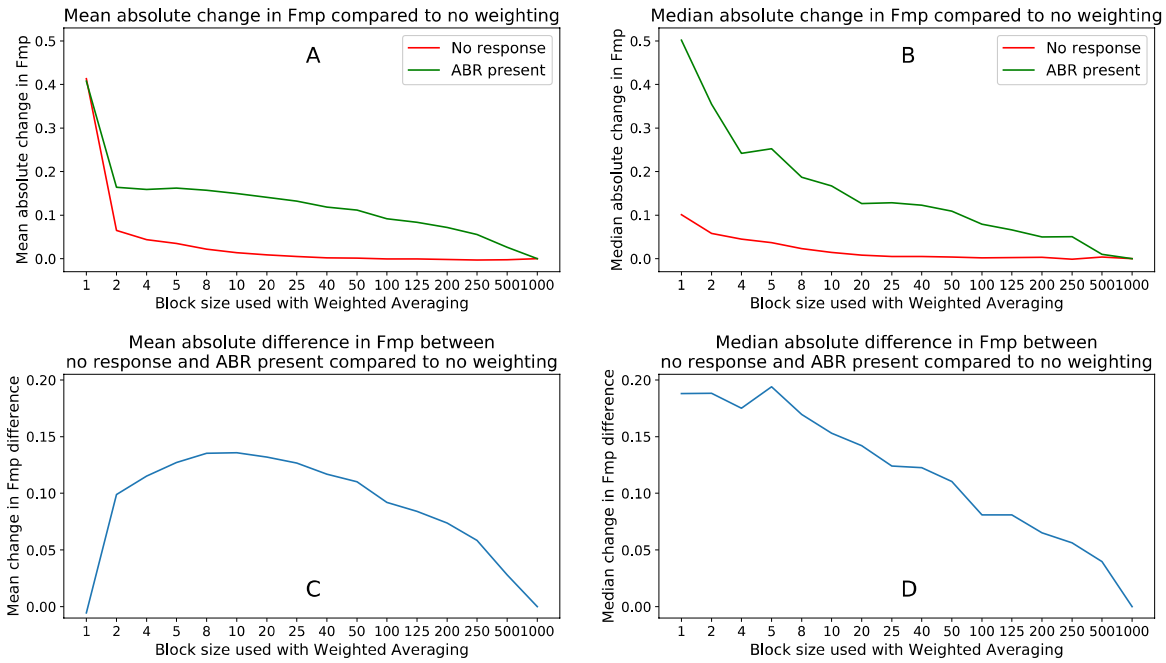

Figure 5-7 Evaluation of the effects of weighted averaging on Fmp values. In all four graphs, the values presented are the absolute difference between the block size in question and a block size of 1,000, i.e. no weighting. Graphs A and C are concerned with mean values, whereas graphs B and D are concerned with median values. Figure reproduced with changes made, in accordance with the CC BY 4.0 license, from McKearney, R. M. et al. (2023) 'Optimising Weighted Averaging for Auditory Brainstem Response Detection', *Biomedical Signal Processing and Control*, 83, p. 104676. Available at: https://doi.org/10.1016/j.bspc.2023.104676.

In Figure 5-7, in graphs A and B, it can be seen that as the block size used with weighted averaging decreased, the mean and median Fmp values of ABR present ensembles tended to increase. It can also be observed that an unintentional side-effect of the weighted averaging procedure using the 'VAR Whole Block' method was an inflation of the Fmp values for 'response absent' ensembles, predominantly affecting small block sizes. This would serve to increase the false positive rate, i.e. the proportion of no-stimulus recordings that are incorrectly determined to contain a response. To evaluate this further, a separate individual evaluation on the effects of weighted averaging on the numerator (evoked potential signal estimate) and denominator (noise estimate) of the Fmp equation is shown in Figure 5-8. A separate evaluation (shown in Appendix D—Figure A 5) using the 'VAR MP' method found this same phenomenon to be present and in fact to a larger extent.

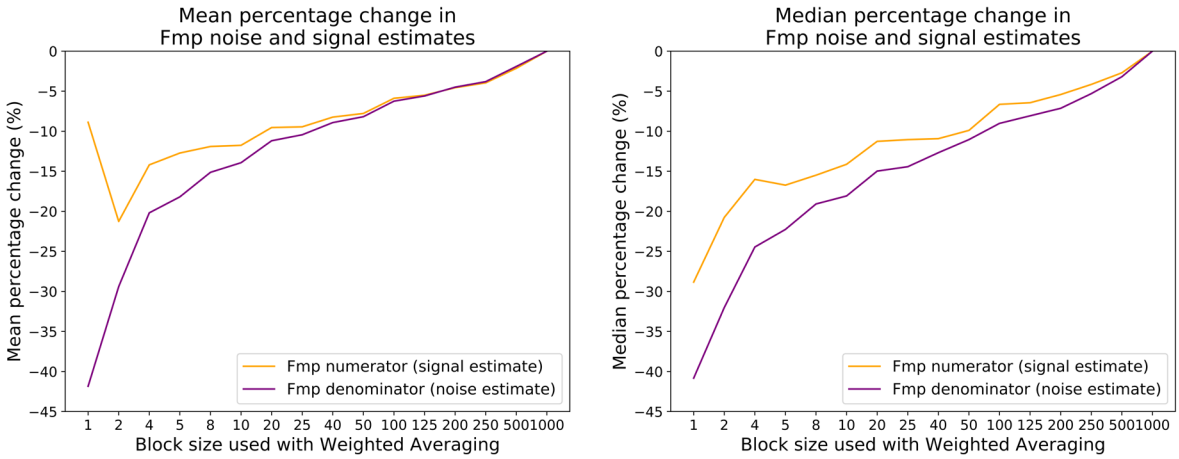

Figure 5-8 Evaluation of the effects of weighted averaging on the numerator (evoked potential signal variance estimate) and denominator (noise variance estimate) of the Fmp equation for 'response absent' data. It can be seen that whilst the mean and median estimate of the variance of the ABR signal decreased with decreasing block size, the mean and median estimates of the variance of the noise decreased by a greater extent, resulting in the inflated Fmp values observed in Figure 5-7 for 'response absent' data.

Figure 5-9 shows the mean change in Fmp value when applying weighted averaging, broken down by the original unweighted Fmp value of the ensemble. The Fmp statistic is expected to follow an *F*-distribution with a mean value of ~1 (Elberling and Don, 1984), as shown in the equation provided by Mood, Graybill and Boes (1974):

$$E[F] = \frac{v_2}{v_2 - 2} \quad for\ v_2 > 2 \tag{5.12}$$

The mean Fmp value obtained empirically from the unweighted no-stimulus data was 0.952, compared to an expected mean value of 1.002 (Equation 5.12). It can be seen from Figure 5-9A

that for ensembles with an unweighted Fmp of less than one, the weighted Fmp value tended to increase; in contrast, for ensembles with an unweighted Fmp of more than one, the weighted Fmp value tended to decrease. As there was a preponderance of low unweighted Fmp value ensembles within the dataset (Figure 5-9B), the mean Fmp value across all ensembles tended to increase when weighted averaging was applied.

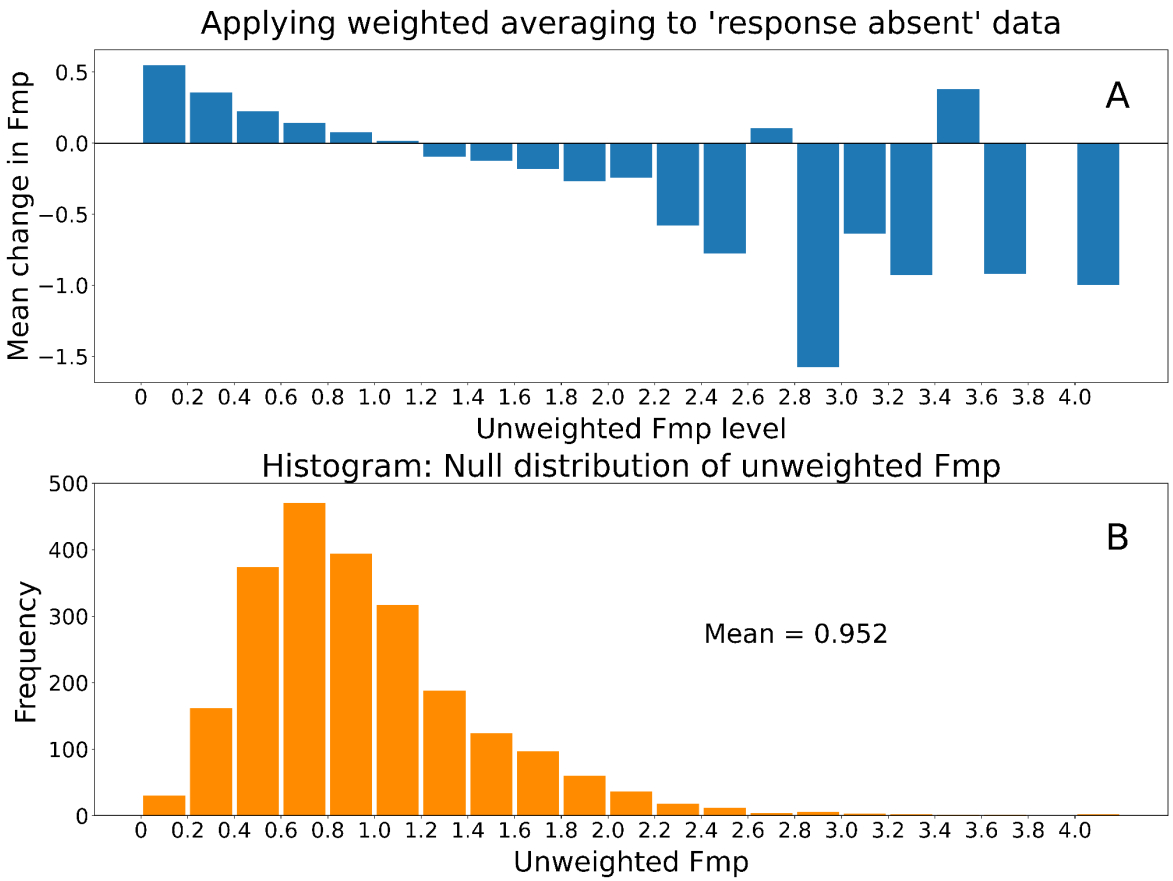

Figure 5-9 Analysis of the null distribution of the unweighted Fmp statistic and the impact of weighted averaging. Graph A shows the mean change in Fmp when applying weighted averaging with 2 epochs-per-block, compared to the original unweighted Fmp value. Graph B shows the null distribution of the unweighted Fmp statistic. Figure reproduced with changes made, in accordance with the CC BY 4.0 license, from McKearney, R. M. et al. (2023) 'Optimising Weighted Averaging for Auditory Brainstem Response Detection', *Biomedical Signal Processing and Control*, 83, p. 104676. Available at: https://doi.org/10.1016/j.bspc.2023.104676.

Figure 5-10 shows the density plot of the unweighted 'response absent' data Fmp statistic compared to the density plot of the closest-matching *F*-distribution with v_2 constrained to be equal to 999 *df* (v_1=7). A one-sample permutation test (using 20,000 permutations) was performed in order to test the null hypothesis that there was no difference between the observed null mean Fmp value and the expected mean value of an *F*-distribution with $v_2 = 999$ *df*

(calculated using Equation 5.12 to be 1.002); the two-sided *p* value was <0.001, meaning that the difference between the observed mean and expected unweighted null Fmp statistic was statistically significantly different. If this result did not occur due to chance, then what was the reason for this difference in Fmp distribution? Is it possible that the assumptions of the *F*-test were violated?

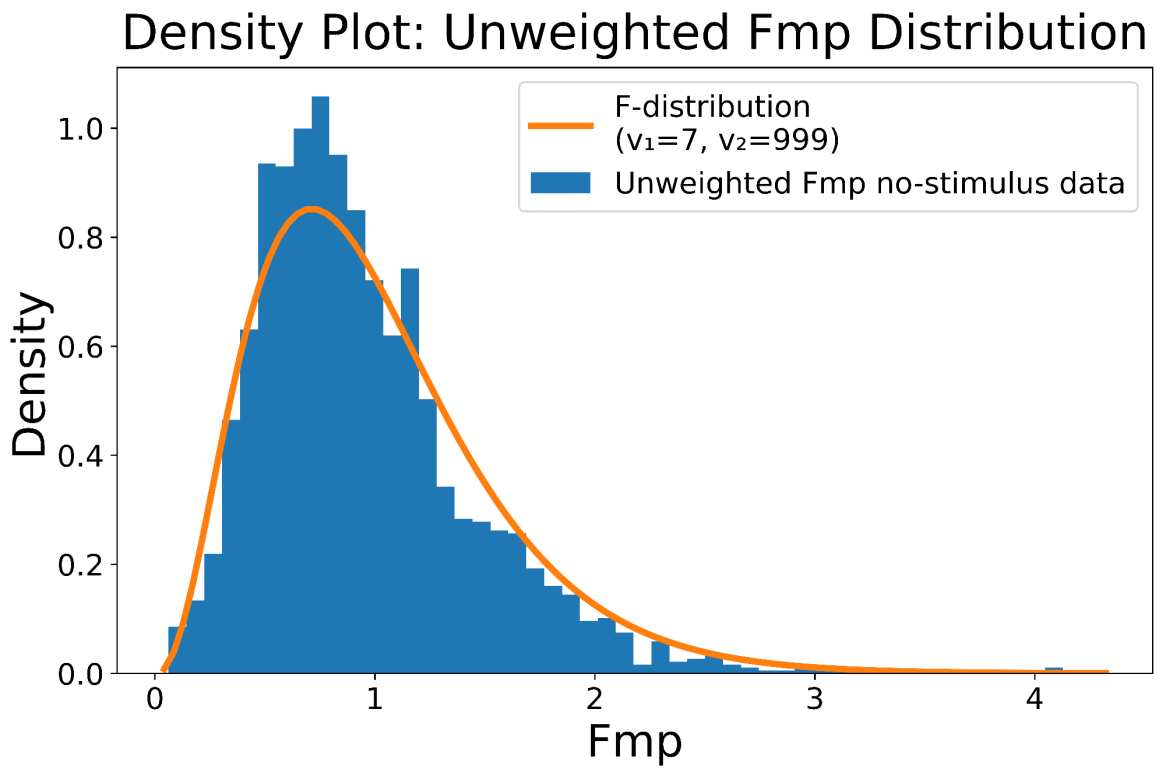

Figure 5-10 Density plot of the unweighted no-stimulus Fmp statistic compared to the closest-fitting *F*-distribution.

The deviation of the null distribution of the Fmp statistic from the expected *F*-distribution may be contributed to by a number of factors, including:

- Independence violations.
- Violation of the assumption of normality.
- An interaction between artefacts, the artefact rejection threshold, and the filter settings.
- Non-stationarity in the data.
- Chance.
- Any combination of the above.
- Other.

The *F*-test of equality of variances is sensitive to non-normality (Pearson, 1931; Box, 1953). The *F*-test also makes the assumption that the two variances in the ratio are independent (Kenny, 1953), and that the samples are randomly selected (Mood, Graybill and Boes, 1974). In a study by

Zimmerman and Zumbo (1992) using simulations to examine the effects on non-independence between sample observations in ANOVA *F*-tests, it was observed that non-independence of sample values between groups led to a significant decrease in type I errors, and an associated increase in the probability of a type II error. This was initially thought to be the potential cause of the lower-than-expected null *F* statistic. Zimmerman and Zumbo (1992) note that violations to the independence assumption 'can have severe consequences for significance testing'.

An investigation into the effects of non-independence was performed, however, it later transpired that the simulated serial correlation was not the cause of the lower-than-expected mean Fmp value. This simulation study is presented in Appendix E. In summary, it was found that by increasing the strength of the low-pass filter applied to randomly generated white Gaussian noise, i.e. increasing dependence between samples, the distribution of the samples became less normally distributed and the mean Fmp value decreased. As it later emerged, the decrease in the null Fmp value observed when increasing the strength of the low-pass filter was not a result of increasing non-independence between samples.

After much deliberation regarding the cause of the lower-than-expected mean Fmp value, my supervisors discussed this finding with a fellow researcher investigating auditory evoked potentials who was able to shed light on the matter. Dr Jaime Undurraga (J. Undurraga, personal communication, 2022) suggested that the observed low mean Fmp value may be due to the length of the analysis window, a caution that Elberling and Don (1984) voice in their original Fsp paper. This effect has also been reported on by the British Society of Audiology (2019c), where the median of the null Fsp statistic was found in a particular study to be noticeably below 1. Elberling and Don (1984) advise that if the analysis window is fixed in length, then the estimated variance of the coherent average will not fully reflect the signal power of frequency components below the level of $1/analysis\ window\ length\ (in\ seconds)\ \ Hz$. Any frequency components present in the coherent average below this frequency cut-off will not be represented fully in the calculation of the variance of the coherent average, i.e. the numerator of the Fmp equation. This effect does not affect the denominator of the Fsp/Fmp equation, resulting in a bias towards low Fmp values in the null condition. The analysis window in the experiment presented in this chapter was 14 ms in length. Frequency components below 71 Hz and above the high-pass filter level of 30 Hz would therefore be partially excluded from contributing to the Fmp numerator but be present in the denominator. The decrease in mean Fmp value produced in the simulations in Appendix E, were therefore due to the variance estimate of the coherent average (the Fmp numerator) not fully reflecting the power of low-frequency signal components, rather than the serial correlation introduced by the autoregressive filter. Figure 5-11 demonstrates the effects of analysis window length on the mean Fmp value, using simulated coloured noise (using the

autoregressive filter as described in Appendix E, with a filter numerator coefficient value of 0.7). The analysis window over which to calculate the Fmp value was applied to the central samples within an ensemble of size 1,000 epochs by 1,000 discrete-time samples (2,000 simulations used). Shorter analysis windows corresponded to a lower Fmp value, due to a decrease in the Fmp numerator (Figure 5-11). It is important that the results presented in this chapter are interpreted in light of the bias present in the Fmp statistic as a result of the fixed analysis window length.

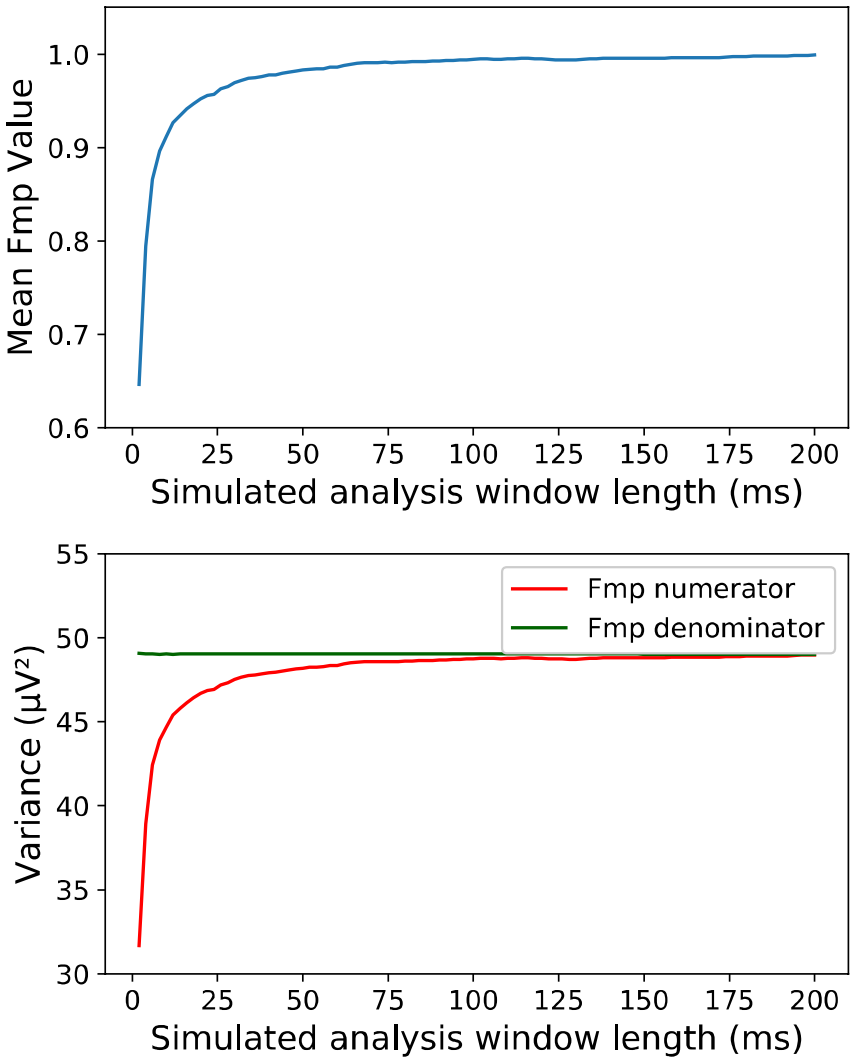

Figure 5-11 Effect of analysis window size on Fmp value. Figure reproduced without changes, in accordance with the CC BY 4.0 license, from McKearney, R. M. et al. (2023) 'Optimising Weighted Averaging for Auditory Brainstem Response Detection', *Biomedical Signal Processing and Control*, 83, p. 104676. Available at: https://doi.org/10.1016/j.bspc.2023.104676.

5.3.2.2 Specificity

A reliable level of specificity across all conditions is critical for an ABR detection method to be deemed reliable for use in clinical applications. Figure 5-12 shows the specificity obtained across a range of block sizes using the Fmp detection method. The significance of an Fmp value is traditionally obtained from a theoretical *F*-distribution with a conservatively selected five *df* (Elberling and Don, 1984; British Society of Audiology, 2019c). Conservative selection of this parameter is hypothesised to be the cause of higher-than-expected specificity levels (Chesnaye *et al.*, 2018), i.e. the observed overly high specificity achieved across block sizes of 2–1,000. It is possible that the lower-than-expected mean Fmp value in the null distribution is also (partially) responsible for the higher-than-expected false positive rates observed. For the block size of one epoch-per-block, the specificity obtained was much lower than expected. Comparing these results to Figure 5-7; the mean Fmp value of the 'response absent' data increased sharply for one epoch-per-block, coinciding with the sharp decrease in specificity observed in Figure 5-12. For all of the other block sizes, the Fmp inflation for no-stimulus data was low and therefore did not appear to significantly affect specificity performance.

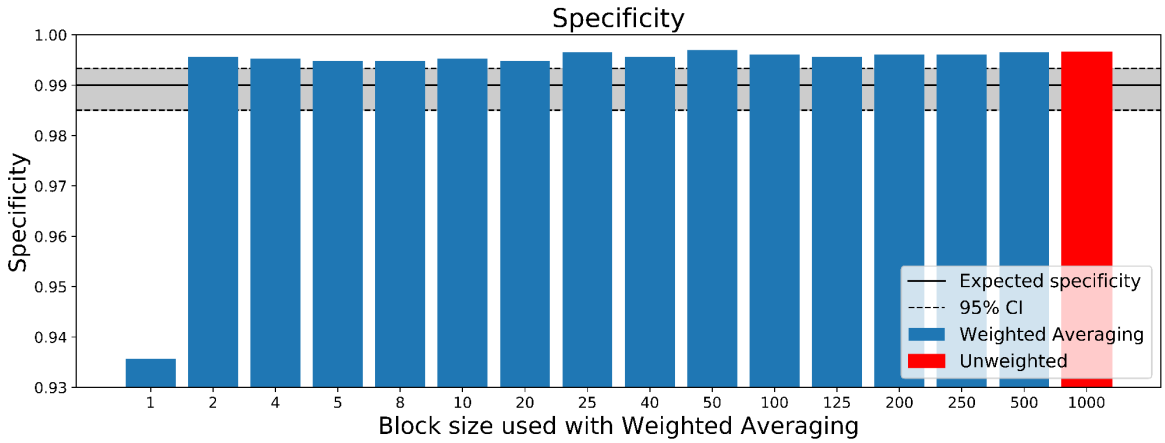

Figure 5-12 Fmp specificity using weighted averaging. Specificity was measured as the proportion of 'response absent' data correctly identified as containing no response. Figure reproduced with minor adaptations, in accordance with the CC BY 4.0 license, from McKearney, R. M. et al. (2023) 'Optimising Weighted Averaging for Auditory Brainstem Response Detection', *Biomedical Signal Processing and Control*, 83, p. 104676. Available at: https://doi.org/10.1016/j.bspc.2023.104676.

5.3.3 Sensitivity

Figure 5-13 shows an evaluation of the sensitivity level achieved per block size used with weighted averaging. It also includes a break-down of performance according to the SNR of the

'response present' data. The sensitivity is presented with the Fmp critical value adjusted to the level which maintained a false positive rate of exactly 0.01. This prevents a detection method where the specificity level is lower from having an unfair advantage (e.g. choosing a detection criterion which produces a specificity of 0 could yield a sensitivity of 1) (Chesnaye *et al.*, 2018). Across all of the 'response present' data, the improvement in sensitivity observed between unweighted averaging and weighted averaging using 10 epochs-per-block was modest (0.410 increasing to 0.467, i.e. a 13.9% relative increase). The low detection rate is in large part attributable to the inclusion of low SNR data. When the ABR present data were stratified into three equal sized SNR groups (low-, mid-, and high-SNR data), the improvement in detection performance using smaller block sizes was less marked for the low- and high-SNR strata. These ensembles likely correspond to data well below or well above the detection criterion. A greater effect was observed in the mid-SNR range as these data reflect the ensembles straddling the detection criterion on the cusp of being detected/not detected. The sensitivity for detecting these mid-SNR 'response present' ensembles increased from 0.257 to 0.368 comparing unweighted averaging to weighted averaging using a bock size of 10 epochs-per-block. This absolute increase in sensitivity of 0.111 corresponds to a relative improvement in detection of 43%. The data for this experiment, but instead using the 'VAR MP' method, are available in Appendix D for comparison. In summary, the graphs for the 'VAR MP' method followed the general trend of those shown in Figure 5-13 (using the 'VAR Whole Block' method) with a greater improvement in sensitivity observed for mid SNR data compared to low- and high-SNR data, however, the peak sensitivity levels were lower when using the 'VAR MP' method. Any observed gain in sensitivity from using a smaller block size being harnessed, relies on being able to control the false positive rate.

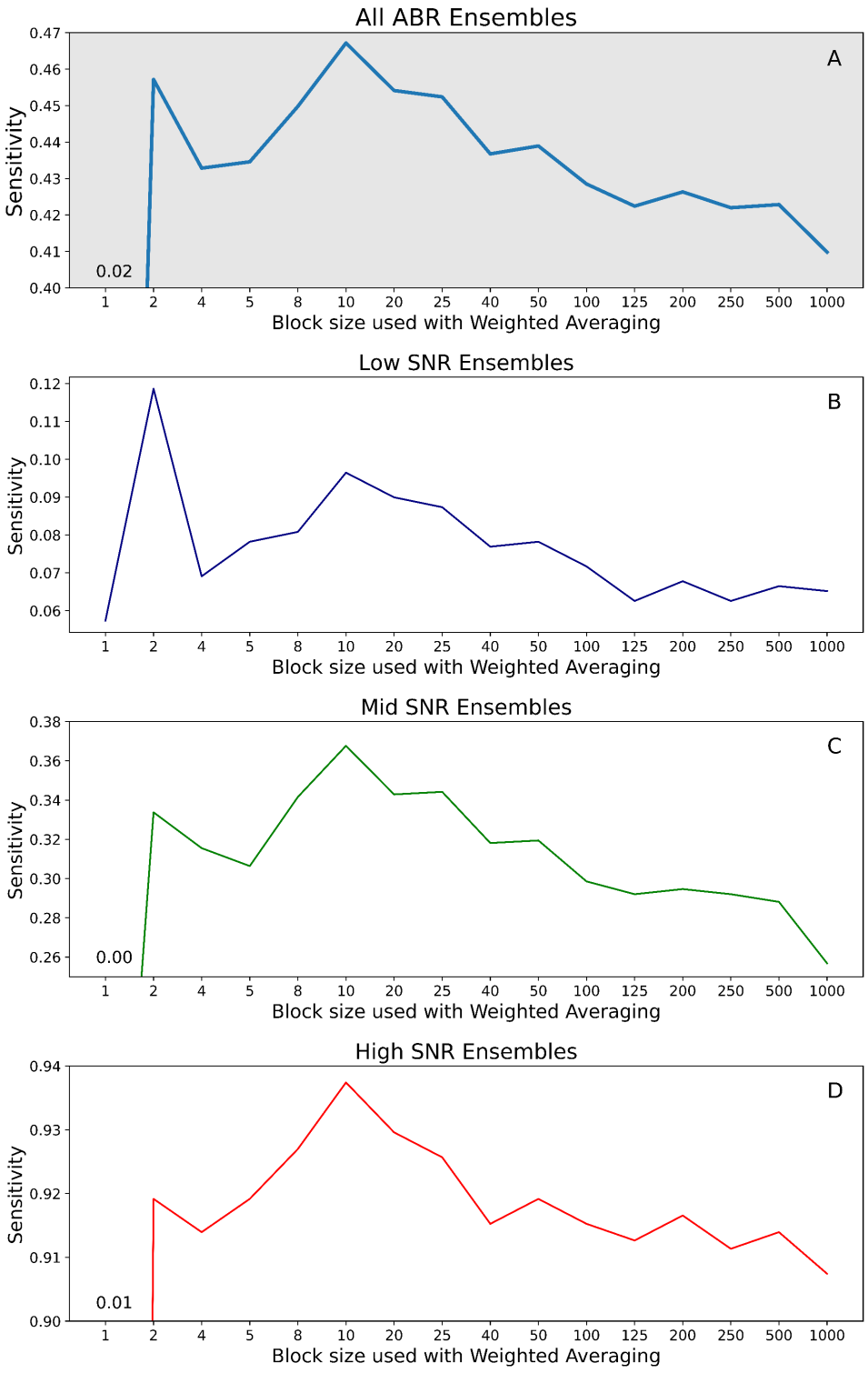

Figure 5-13 Sensitivity achieved across different block sizes. In order to assess the level of sensitivity fairly, the Fmp critical value was adjusted to that which achieved the desired false positive rate (0.01) exactly. Plot A shows the sensitivity level across block sizes as the proportion of all of the 'response present' ensembles correctly detected. For graphs A, B and C, the 'response present' data were stratified into three evenly split groups of low- (< -32 dB), mid- (-32 to-27 dB), and high-SNR

(> -27 dB) 'response present' data. The sensitivity was then calculated for each portion of the 'response present' data.

5.3.4 Controlling the False Positive Rate

Data processing techniques such as weighted averaging may alter the properties of the data (Lütkenhöner, Hoke and Pantev, 1985) and potentially alter the performance of statistical detection methods. Figure 5-14 shows how weighted averaging (in combination with the data processing parameters used) can alter the Fmp null probability distribution. It could be that this is a result of the Fmp bias introduced by the analysis window length used. Weighted averaging is expected to affect low-frequency components in the EEG data the most (Hoke *et al.*, 1984). This effect would vary by block size, with the effect of the Fmp bias therefore also varying. In order for the benefits of weighted averaging to be harnessed, the false positive rate must be controlled so that the detection method performs stably and reliably across a variety of data. One such method of controlling the false positive rate is through the use of the bootstrap technique described in Section 3.3 (The Bootstrap Technique).

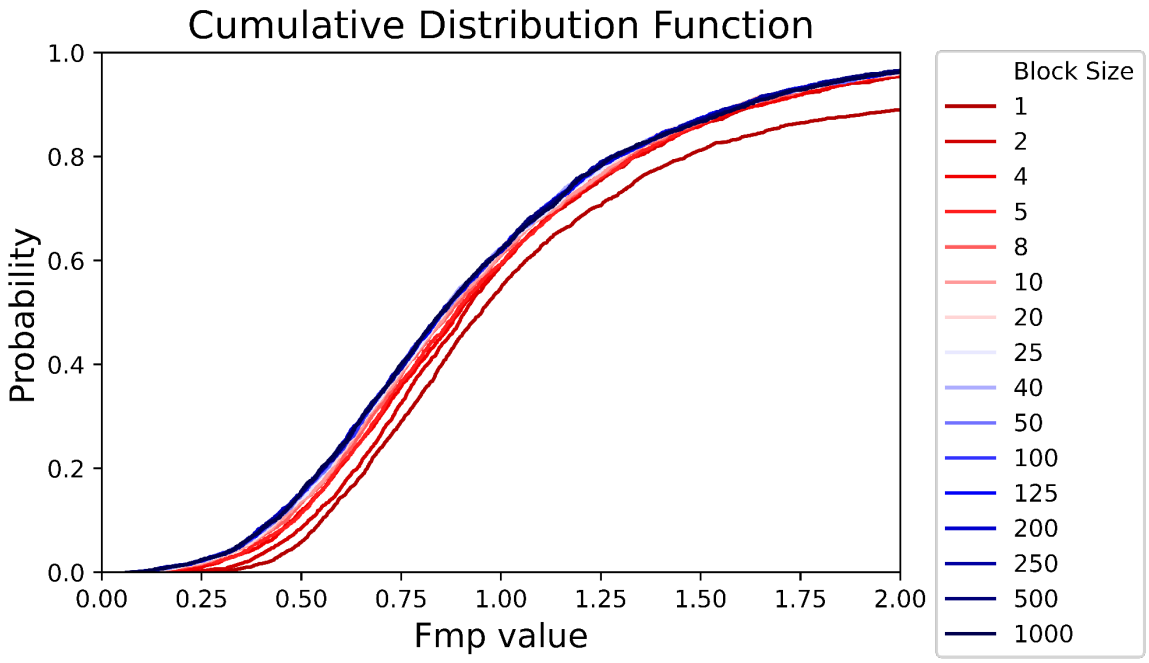

Figure 5-14 Weighted averaging can alter the null probability distribution. For low block sizes, a right-shift in the null probability distribution was observed, corresponding to an inflation in the Fmp values of the 'response absent' data.

Figure 5-15 shows how the original bootstrap technique (Lv, Simpson and Bell, 2007) performed at stabilising the false positive rate. Using the bootstrap technique, the specificities recorded were consistently within the expected 95% CI across block sizes, even when no weighted averaging was

applied (i.e. unweighted averaging). The binomial proportion 95% CI was calculated using the Wilson score interval method (Wilson, 1927).

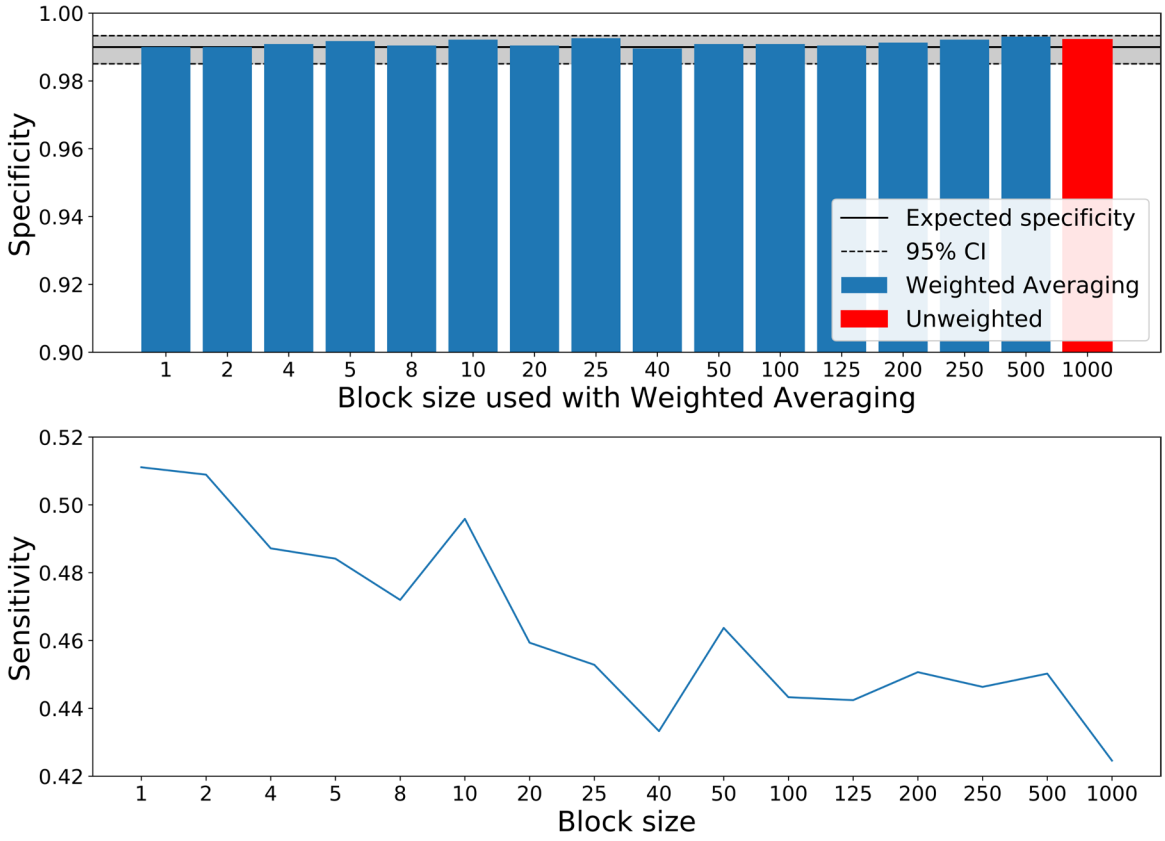

Figure 5-15 Controlling the false positive rate using the bootstrap. The top graph shows the specificity achieved using the Fmp statistic combined with weighted averaging and the bootstrap technique. The bottom graph shows the sensitivity achieved using this method, with the critical value adjusted to give a false positive rate of exactly 0.01. Figure reproduced with changes made, in accordance with the CC BY 4.0 license, from McKearney, R. M. et al. (2023) 'Optimising Weighted Averaging for Auditory Brainstem Response Detection', *Biomedical Signal Processing and Control*, 83, p. 104676. Available at: https://doi.org/10.1016/j.bspc.2023.104676.

Using the bootstrap technique allowed a higher detection rate to be achieved using smaller block sizes, without incurring an increase in the false positive rate.

5.3.5 Analysis of Simulated Stationary Data

Figure 5-16 shows the distribution of a measure of the stationarity of the noise within each 'response absent' ensemble. A large portion of the ensembles were mostly stationary. However, there are some ensembles which displayed a high degree of non-stationarity.

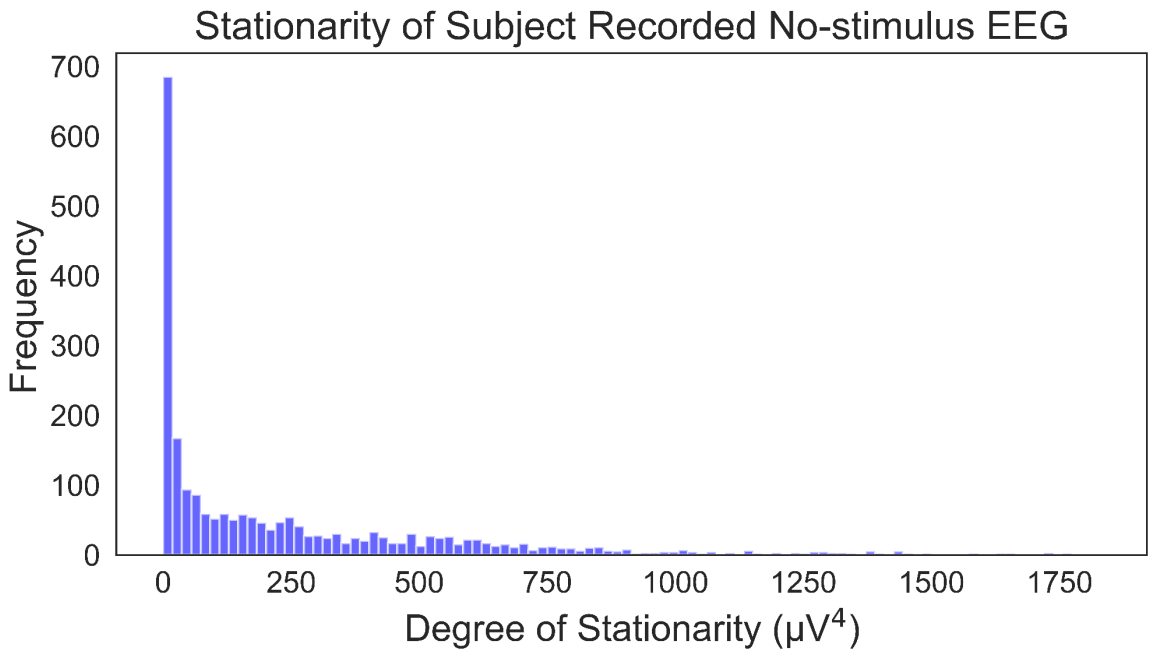

Figure 5-16 A histogram presenting the stationarity of the subject recorded no-stimulus EEG ensembles. The degree of stationarity was calculated as the variance of the variance of each of the 1,000 recording epochs in each ensemble. Data where the noise variance in each recording epoch was identical would be expected to have a value of zero.

Weighted averaging confers a benefit over unweighted averaging only when the data are non-stationary. It is important that detection methods combined with weighted averaging work well also when the background noise is stationary. Figure 5-16 shows an analysis of weighted averaging using simulated stationary data (see 5.2.7). Both methods of estimating the variance of the noise for applying weighted averaging performed increasingly worse using smaller block sizes, however, the 'VAR MP' method performed worse than the 'VAR Whole Block' method for these stationary data. These data highlight the danger of selecting too low a block-size where the estimate of the noise level in each block become inaccurate. Note that this result may be influenced by the length of the Fmp analysis window used.

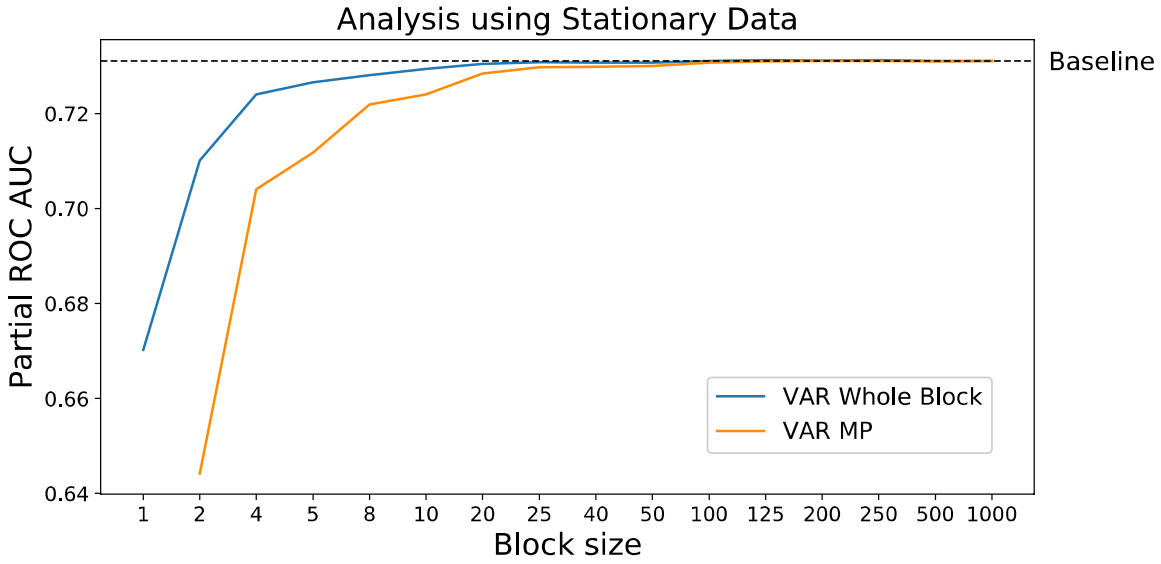

Figure 5-17 Analysis using simulated stationary data where the noise had equal variance across all recording epochs in the ensemble. Unlike in Figure 5-4 where weighted averaging was evaluated on a dataset including non-stationary data, there is no increase in partial ROC AUC observed, only a decrease in performance when too low a block size was selected. Weighted averaging can therefore be harmful to detection performance for stationary data if too low a block size is chosen.

5.3.6 Machine Learning—Feature Comparison

Relating the work of the current study to the research presented in Chapter 4, weighted averaging may be used as a feature extraction technique to provide input features to the machine learning model which have a higher SNR compared to when using unweighted averaging. An additional research question was therefore proposed: providing the coherently weighted average to the stacked ensemble as an input feature will lead to improved ABR detection performance, compared to when using the unweighted coherent average. To compare performance when using the different input features, the stacked ensemble presented in Chapter 4 was trained on two versions of the training set: either the original features including the unweighted average, or the original features with the unweighted average replaced with the weighted average were used. Note that the other input variables, such as all of the input variables to the random forest branch of the stacked ensemble, were the same for both feature sets. Using optimised hyperparameters the stacked ensemble was trained on each training feature set and then evaluated on the test set data (containing the same extracted features), repeated 50 times, to produce Figure 5-18. No significant difference in test set performance was found between the two feature sets (Mann-Whitney U test; U=1180, p=0.32 two-tailed). This may reflect the incremental benefits in residual noise reduction using weighted averaging compared to unweighted averaging (Figure 5-6).

Moreover, these incremental benefits may be eclipsed by the importance of the multiple other input features used by the stacked ensemble, rendering a slight improvement in the quality of this one input feature redundant.

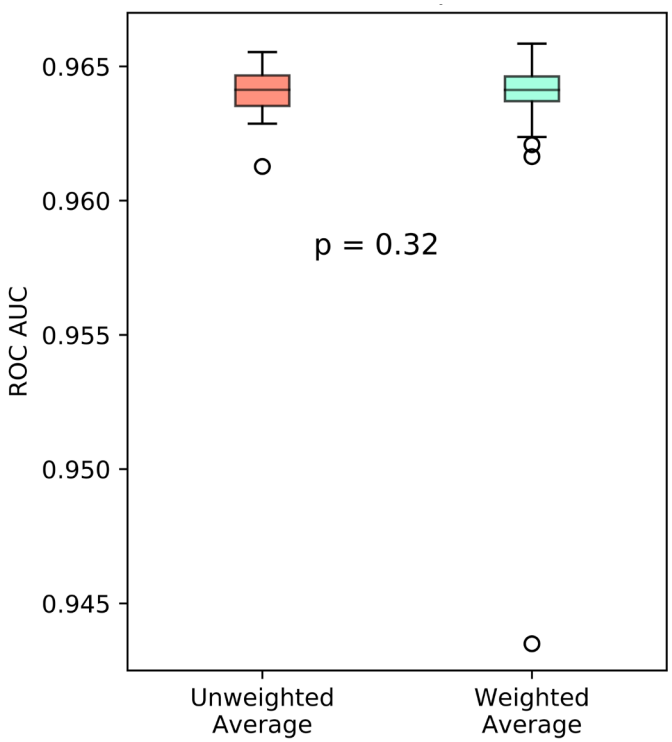

Figure 5-18 Feature comparison. The stacked ensemble was trained on each feature set containing either the unweighted average or weighted average (amongst the many other input features used by the stacked ensemble) and evaluated on the test set over 50 iterations.

5.4 Discussion

Weighted averaging is an effective technique for reducing the residual noise level present in the coherently averaged waveform (Hoke *et al.*, 1984; Elberling and Wahlgreen, 1985). If the noise level within each recording epoch can be accurately estimated, weighted averaging using inverse-variance weighting is known to minimise the variance present in the evoked potential signal estimate (Hartung, Knapp and Sinha, 2008).

5.4.1 Observed Bias in the Fmp Statistic

Data pre-processing algorithms such as weighted averaging can alter the properties of the data (Lütkenhöner, Hoke and Pantev, 1985), which may have unintended effects upon the performance of ABR detection methods. Figure 5-11 shows how a short-length analysis window led to lower-than-expected mean Fmp values. This is due to the exclusion of signal power for

frequencies below the cut-off imposed by the finite-length analysis window (Elberling and Don, 1984). Increasing the length of the analysis window is one potential solution. However, increasing the length of the analysis window beyond the duration of the evoked potential being measured, has the potential to reduce the SNR obtained. Elberling and Don (1984) advocate 'appropriate high-pass filtering', i.e. increasing the level of the high-pass filter so that it is equal to or greater than the low-frequency component cut-off introduced by the finite analysis window length. This has the limitation of potentially excluding low-frequency components of the ABR signal. Note that the high-pass filter setting used in this study was 30 Hz, based on the recommendations of the BSA for ABR testing in infants (British Society of Audiology, 2019c). In order to investigate the effects of the high-pass filter setting, a simulation was performed, reanalysing the data used in the present study, but with the high-pass filter settings raised from 30 to 100 Hz. This avoids low-frequency signal power from being excluded from the Fmp numerator due to the 14 ms window length used. The results of this simulation are presented in Appendix F. In summary, the results showed that, in terms of ABR detection and residual noise reduction, the 'VAR Whole Block' method still performed more effectively than the 'VAR MP' method, although the differences in performance between methods were less noticeable than in Figure 5-4. This is likely because raising the high-pass filter noticeably reduced the background noise levels, minimising the impact that weighted averaging can have. The mean null Fmp value in the dataset was 1.02 (expected value 1.002). The inflation in the 'response absent' Fmp statistic observed when weighted averaging was applied (Figure 5-7) may have been as a result of the low mean null Fmp statistic and caused by the effect of the finite analysis window length.

When weighted averaging is applied, this will serve to reduce the frequency components most present in the noise, i.e. low frequencies for EEG data. By reducing the low-frequency noise content, weighted averaging would minimise the effect of the bias introduced by the analysis window length, thereby increasing the null Fmp value. Whilst this explanation is plausible, it should be noted that whilst raising the high-pass filter from 30 to 100 Hz much reduced the 'response absent' Fmp inflation, the median Fmp value was still significantly raised when applying weighted averaging with a block size of one (Figure A 10). Further investigation into the effects of analysis window length, filter settings and weighted averaging parameters is warranted.

Of note, the ABR detection performance was improved when raising the high-pass filter from 30 to 100 Hz (Figure A 9). This is likely because the SNR of the data was increased by excluding the low-frequency power of the background noise. It should be noted, however, that this result is based on simulations using one single ABR template (spectral content will vary depending on the ABR template used). Further work is required to investigate the optimal filter settings, using clinical data recorded from a large number of individuals.

As a note of caution, a finite analysis window length which (partially) excludes low-frequency components in the Fmp numerator may have the potential to reduce ABR detection performance. This is because, as well as underestimating low-frequency signal power of the averaged background noise (which contributes to the Fmp numerator), it may also underestimate the low-frequency power of the evoked potential signal (also contributing to the Fmp numerator). A short analysis window length, coupled with a high-pass filter setting which partially excludes low-frequency signal power, will underestimate the averaged background noise level in the Fmp numerator as well as, possibly, the evoked potential for 'response present' data, decreasing the value of the Fmp statistic. The analysis window Fmp bias is therefore hypothesised to potentially have a double-reduction effect on the Fmp value of 'response present' data, where the Fmp value may be reduced by a greater factor than that of the 'response absent' data, reducing ABR detection performance. Increasing the length of the analysis window will allow more low frequency evoked potential signal power into the Fmp numerator variance estimate. This will only increase the SNR *overall* if the expansion of the analysis window length does not diminish the average power of the evoked potential signal by including latencies which capture little ABR signal power, i.e. it is possible that whilst increasing the analysis window length may introduce greater low-frequency signal power, the benefit to the overall SNR may be outweighed by the analysis window being expanded to include regions of low SNR.

An alternative method to overcoming the bias in the Fmp statistic is to use the bootstrap technique. This has been shown to control the false positive rate (Figure 5-15) but may not overcome the effect of the Fmp bias on ABR detection performance. The bootstrap method may also be used in combination with the above recommendations. Figure 5-7 shows that whilst the Fmp value is boosted by weighted averaging for 'response present' ensembles, it can also be inflated for 'response absent' ensembles if too low a block size is used, leading to a raised false positive rate. Methods to control the false positive rate are required to harness any boost in detection offered by weighted averaging and to ensure that a stable level of specificity is achieved. The bootstrap technique (Lv, Simpson and Bell, 2007; Chesnaye *et al.*, 2018; Chesnaye, 2019) (Section 3.3) was found to successfully control the false positive rate across all of the block sizes evaluated (Figure 5-15), allowing even smaller block sizes to be used, achieving an increase in detection performance without an increase in the false positive rate.

A high-pass filter setting of 30 Hz was chosen, based on the recommendation of the BSA guidelines for ABR testing in babies (British Society of Audiology, 2019c). Whilst a high-pass filter setting of 100 Hz was found later to be more effective than 30 Hz (Appendix F), the main data presented in this chapter reflect the 30 Hz condition. This is because, whilst performance was not as good, these data reflect more accurately what may be being currently observed in clinical

practice. These data also serve to warn clinicians on the combined effects of the Fmp analysis window length and the high-pass filter setting, especially when applying weighted averaging. This study highlights the need for further research optimising these parameters, with some additional data using a high-pass filter setting of 100 Hz presented in Appendix F. The recommendations in Section 5.4.3 regarding optimisation of the block size parameter are therefore based on the originally selected 30 Hz high-pass filter setting.

5.4.2 Noise level Estimation Methods

Due to the limited number of (independent) discrete-time sample points within each recording epoch it is typically not possible to accurately estimate the noise level in order to compute an accurate weight for each recording epoch. This has led to the approach of weighting groups of epochs together in blocks, rather than weighting them individually. Smaller block sizes allow the weights to be applied more precisely. However, if the block size is too small, the variance of the noise cannot be accurately estimated, and the weights calculated will not accurately reflect the changing noise levels over time within the EEG recording. A trade-off is therefore sought. This study explored two methods of estimating the variance of the background noise: calculating the variance down multiple points across recording epochs (Elberling and Wahlgreen, 1985; Martin *et al.*, 1994), and calculating the variance of all samples within a block. Figure 5-4 shows how the 'VAR Whole Block' method was able to achieve a higher partial ROC AUC score, compared to the 'VAR MP' method; a statistically significant difference between the two methods was observed for block sizes of 2-to-10 epochs-per-block. This suggests that the 'VAR Whole Block' method was able to estimate the variance of the noise more accurately within each block, allowing more precise weighting using smaller block sizes to be applied. This is consistent with the findings presented in Figure A 3. The 'VAR Whole Block' method increases the degrees of freedom present in the variance estimate, compared to the 'VAR MP' method. However, the 'VAR Whole Block' method has the limitation of introducing a bias factor in the estimate due to the variance estimate being affected by the presence of the evoked potential signal. This type of bias is expected to be of little significance when analysing low SNR signals such as the ABR (Sörnmo and Laguna, 2005). Having said this, selecting the VAR MP method may still be a reasonable option given the relatively small amount of benefit conferred by the VAR Whole Block method and the potential for untoward effects due to the bias in the variance estimate. The accuracy of the noise estimate may be improved further by expanding the analysis window length (if possible) to include more sample points (note that varying the analysis window length may affect the SNR).

5.4.3 Optimising the Block Size Parameter

Having assessed the optimal method for estimating the variance of the noise within each block, the optimal block size was subsequently analysed. A block size of 250 epochs-per block was recommended initially by Elberling and Wahlgreen (1985) using the single point method of estimating the noise level, with the acknowledgement that a lower value for this parameter would likely be more effective yet. Subsequent research showed that a block size of 32 epochs-per block reduced the residual noise within the average the most, estimating the noise level from eight points within each recording epoch (Don and Elberling, 1994). This was the lowest value for the block size parameter evaluated by Don and Elberling (1994), with the indication that even lower values may improve performance further. Indeed, using a form of iterative weighted averaging, Riedel, Granzow and Kollmeier (2001) found a block size of four epochs-per-block to be most effective. However, due the small number of simulated recordings used in that particular analysis, interpretation of the optimal block size is challenging.

Considering the data presented in the Results section, a reasonable best compromise for the block size parameter was found to be approximately 10–20 epochs-per-block. If a block size lower than this is used, performance may decline sharply as the accuracy of the noise level estimate decreases. Whilst median results (residual noise and Fmp levels) improved with lower block sizes, the mean values declined (Figure 5-7). This indicates that whilst a lower block size that optimises median residual noise reduction may be beneficial in the majority of cases, it will result in weighted averaging having a serious detrimental effect on a minority of recordings (increasing the residual noise levels). This increase in residual noise levels when using weighed averaging with a low block size was also observed by Riedel, Granzow and Kollmeier (2001). The optimal block size for clinical evaluation should avoid this drop-off in mean performance and probably therefore err on the side of caution, leading to the recommendation of selecting a slightly larger block size. The decrease in performance caused by cautiously selecting a slightly larger block size than may be optimal, likely outweighs the risks of selecting too low a block size. There is also the risk that the presented results may not tally exactly with those that may be obtained from other sources of EEG data using different recording parameters (e.g. sampling rate and filter settings). Performance too, depends on the degree of non-stationarity of the data. Selecting a slightly larger block size would provide leeway for any variation that may be observed in new data, avoiding the steep drop-off in detection performance observed when too low a block size was selected (Figure 5-4). Based on this and the evidence provided using the present dataset, a block size of 20 is recommended. Using the ‘VAR Whole Block’ method, 20 epochs-per-block produced no sizeable decrease in performance using simulated stationary data, whereas lower block sizes caused an increasing drop-off in performance. For recording settings which may limit the degrees of

freedom of the data (i.e. there being fewer independent samples within the analysis window), an even higher value block size parameter would be recommended in order to allow accurate weights to be computed. Note that these recommendations are based on the recording parameters used in the study, including filter settings based on British Society of Audiology (2019c) recommendations, and the chosen analysis window length (14 ms), which contributed to a bias in the Fmp numerator. It should be noted, that when repeating the analysis using a high-pass filter set to 100 Hz, the optimal block size was one epoch-per-block using the 'VAR Whole Block' method (Appendix F). Further analysis is provided in Appendix F where different filter settings are used to avoid the bias in the Fmp statistic. Raising the high-pass filter to 100Hz, reduced the Fmp inflation for the 'response absent' data.

5.4.4 Limitations and Ideas for Future Work

Whilst the 'response absent' data used in this study is recorded from subjects, the 'response present' data are simulated through the addition of an ABR template (derived from only one recording) to each recording epoch. Whilst the noise and response signal in evoked potential recordings are assumed to be additive (Wong and Bickford, 1980; Elberling and Don, 1984), the results obtained based on simulated data may not extrapolate wholly to subject recorded data. Simulation aside, it is also possible that the results based on subject recorded 'response absent' EEG used in this study may not generalise to other sources of EEG, especially if the recording settings vary to those used in the present study. The frequency spectrum of the EEG, independence between samples, independence between epochs, and Fmp analysis window length, may vary between datasets, and so it is possible that these the recommendations based on the findings presented may not be optimal across all recording setups and patient groups. A cautious approach to the use of processing methods such as weighted averaging is advisable, avoiding very low block sizes which can cause a steep drop in detection performance. The steep drop-off in performance observed for low block sizes may occur at a different block size if the characteristics of the EEG differ. Data factors which may warrant higher levels of caution in the choice of block size include:

- A lower number of sample points in the block of epochs from which the variance of the noise is estimated.
- A higher sampling rate, meaning that there will be greater correlation between samples within recording epochs which serves to reduce the number of independent samples in a noise estimate (assuming the overall number of samples in the analysis window remains the same).

- Spectral content of the EEG; EEG noise containing ‘a narrow band of dominant frequencies’ (Elberling and Don, 1984) will contain fewer independent samples from which to accurately estimate the variance of the noise in the block.

A conservative choice of block size (erring on the side of choosing a larger value for this parameter) may not provide optimal detection performance but will help to avoid the significant untoward effect of selecting too low a block size which was observed. Whilst very low block sizes were found to be suboptimal, the peak performance by block size is quite level across a range of block sizes (Figure 5-6—residual noise reduction), providing a wide choice of reasonable block size parameter values. It should be noted that much of the analysis was performed using an Fmp analysis window of 14 ms as described in the Methods section. This introduced a bias to the Fmp statistic. This works highlights the importance of being cautious when considering the Fmp analysis window length and filter settings when applying weighted averaging to ABR data. Whilst additional data using a 100 Hz high-pass filter setting was found to improve ABR detection (Appendix F), the results are based on simulation using only one single ABR template. Further work is required to evaluate the combined effects of Fmp analysis window length, filter settings and weighted averaging parameters using a large database of subject recorded data before a definitive recommendation of parameter settings can be made.

The Fsp/Fmp statistic has been observed to have a lower-than-expected false positive rate (Figure 5-12—block sizes 2–1,000) (Chesnaye *et al.*, 2018; Chesnaye, 2019; McKearney *et al.*, 2022). This has been attributed to the conservative choice of five *df* applied to the Fmp numerator, when calculating the *p* value for a given Fmp statistic, when in practice the number of degrees of freedom within the coherent average is typically greater than this (Elberling and Don, 1984; Chesnaye *et al.*, 2018). It is possible that this may also be partially contributed to by the low mean Fmp values caused by the analysis window length. Further work is recommended whereby the high-pass filter setting and analysis window length is varied, and the mean null Fmp statistic is measured as the dependant variable. This will help to determine the relative contributions of these two factors in causing the lower-than-expected false positive rate.

The current work shows that optimised weighted averaging may significantly increase detection performance, however, the optimal recommendation for block size likely needs to be determined for specific recording setups. It is not possible to provide a single recommendation for optimal block size that will cover all eventualities. Further work is recommended to explore the combined effects of varying recording parameters, the weighted averaging noise estimation method, and the block size used with weighted averaging on the ABR detection performance levels prior to implementing any changes to clinical evoked potential equipment software on the basis of the

research in this chapter. The Fmp statistical detection method was chosen for evaluation in the current study due to the ubiquitous use of this detection method and its relation the Fsp in clinical auditory evoked potential recording software. These methods have been found to perform less well than Hotelling's T^2 test and the q-sample uniform scores test and its modifications (Chesnaye *et al.*, 2018). Future work may explore the effects of combining weighted averaging with these alternative ABR detection methods.

5.5 Conclusions

Weighted averaging can be used to improve the SNR within the averaged waveform as well as improve ABR detection using the Fmp. Weighted averaging using a block size of 10 led to a 13.9% relative increase in ABR detection rate using the Fmp (corrected to maintain the same false positive rate), compared to conventional unweighted coherent averaging. Weighting blocks of epochs inversely to the variance of the noise using the 'Whole Block' method proved more effective than the 'VAR MP' method. Block size is an important parameter in weighting averaging. It must neither be so high such that weights are applied imprecisely, nor be too low—causing the noise level and therefore the weights to be inaccurately estimated. A block-size of 20 epochs-per-block achieved near-optimal residual noise reduction for the dataset used, whilst providing some leeway for differences in EEG characteristics, and avoiding the roll-off in detection performance observed when using very low block sizes. This choice may be influenced by recording settings and if in doubt, it is safer to select a slightly larger block size than a smaller one. The bootstrap technique may be used to control the false positive rate, allowing the benefits in ABR detection from weighted averaging to be harnessed whilst preserving a stable level of specificity. The chosen Fmp analysis window length was found to bias the Fmp statistic and in turn the performance of weighted averaging. This study highlights the caution required when selecting Fmp analysis window length and weighted averaging parameters. Raising the high-pass filter from 30 Hz to 100 Hz was found to minimise this issue and improved ABR detection further yet. However, further research is required using a large amount of clinical data to investigate the optimal filter settings, analysis window length, and weighted averaging parameters prior to recommending any definitive changes to current clinical practice.

Chapter 6 Automated Analysis of the Diagnostic ABR using Machine Learning

6.1 Introduction

6.1.1 Literature Review

6.1.1.1 Clinical Context and Potential Research Impact

The term 'diagnostic ABR' refers here to the ABR recorded using a suprathreshold stimulus in order to evaluate the function of the auditory nerve and auditory brainstem structures (rather than for the purpose of hearing threshold estimation). The application of the ABR for this purpose may also be referred to as the 'neurological ABR' (British Society of Audiology, 2019b). When recorded at a suprathreshold level, the entire morphology of the ABR waveform may be observed, with all of the key components visible (Figure 6-1).

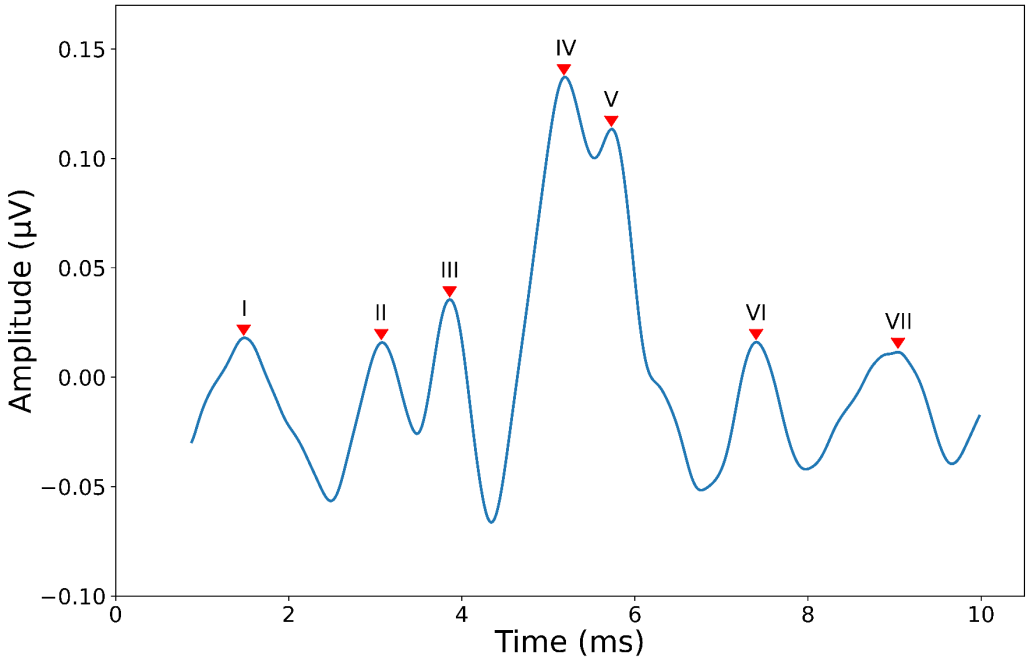

Figure 6-1 The neurological ABR waveform. Waves I–VII are labelled in accordance with the Roman numeral convention provided by Jewett, Romano and Williston (1970).

6.1.1.1.1 Neural Generators of the ABR

The waves of the ABR are generated by structures along the auditory nervous system as bioelectrical activity is propagated along the pathway, following the delivery of an auditory

stimulus. These electrical potentials arise from groups of neurones, e.g. within the brainstem nuclei, which are aligned, activated synchronously, and therefore generate a large-enough electrical dipole to be recordable from surface electrodes on the scalp (Atcherson, 2012). The neural generators of the ABR waves may be attributed to structures within the auditory nervous system pathway. Whilst there is reasonable confidence regarding the anatomical contributors to the early components of the ABR (waves I and II), this is not the case for the later ABR components (Hall, 2007). The reason for this is that the auditory brainstem pathway is complexly interconnected, with both ipsilateral and contralateral pathways, resulting in the conclusion that the generation of the later ABR components is likely contributed to by multiple anatomical sources (Atcherson, 2012). A summary of the likely neural generators of the ABR is provided in Table 6-1. See Figure 2-2 for a diagram of the auditory nervous system anatomy.

Table 6-1 The neural generators of the ABR. The information summarised in this table is provided by Møller (2006) as outlined by Atcherson (2012).

ABR Component	Neural Generator
Wave I	Distal auditory nerve
Wave II	Proximal auditory nerve
Wave III	Cochlear nucleus
Wave IV	Midline brainstem structures including the superior olivary complex
Wave V	Termination of lateral lemniscus with contralateral inferior colliculus

6.1.1.1.2 Diagnostic Uses of the ABR

The diagnostic ABR provides a functional assessment of the auditory nerve and auditory brainstem structures. Pathologies affecting these structures may impact upon the morphology of the ABR, e.g. the presence of the ABR waves, their latency and their amplitudes. A previously common application of the diagnostic ABR was in the detection of acoustic neuromas (Selters and Brackmann, 1977); these are benign tumours which grow on the vestibulocochlear nerve. Tumours affecting the auditory nerve can cause a detectible shift in the latencies of the ABR

waves which are contributed to by neural generators proximal to the site of the lesion (Selters and Brackmann, 1977). Whilst largely superseded by the increasing use of MRI scans, the diagnostic ABR still offers utility in certain clinical situations (Montaguti *et al.*, 2007): in areas where MRI scanners are not readily available (Montaguti *et al.*, 2007), as a screening tool to guide the use of limited MRI resources (Montaguti *et al.*, 2007), and in patients for whom MRI scanning may be contraindicated/not possible, e.g. due to metallic implants, claustrophobia, or severe obesity (Fortnum *et al.*, 2009).

The ABR is recommended to be used as part of the diagnostic test battery in the assessment of possible auditory neuropathy spectrum disorder (ANSD) in young infants (British Society of Audiology, 2019a). ANSD is a hearing disorder which is defined by an absent or grossly abnormal ABR morphology with present cochlear microphonic recordings and/or otoacoustic emissions (British Society of Audiology, 2019a). This pattern of results indicate normal outer hair cell function accompanied by a lack of neural synchrony (Madden *et al.*, 2002).

As well as for the above two conditions, the ABR may be used in the investigation of a wide range of retrocochlear pathologies affecting the functional integrity of the auditory nervous system (British Society of Audiology, 2019b). Some examples of these include disorders of the central nervous system (CNS) as summarised by Hall (2007): neurosarcoidosis (Souliere *et al.*, 1991), cerebellar ataxia (Pal *et al.*, 1995), Cogan's syndrome (Benitez *et al.*, 1990), and CNS miliary tuberculosis (Stach, Westerberg and Roberson Jr, 1998).

6.1.1.1.3 Surgical Monitoring

A further use case for automated ABR analysis is in the field of intraoperative monitoring. Some surgical procedures such as acoustic neuroma resection may be liable to cause damage to the auditory nerve. The ABR acts as a useful neuromonitoring tool, allowing the surgeon to receive direct feedback regarding the function of the auditory nerve during the procedure in order to help preserve hearing function (Hummel *et al.*, 2016). The latency of ABR wave V may be used to monitor auditory function intraoperatively, with an increase in the wave V latency indicating neurological dysfunction (Hall, 2007). Surgical procedures can be lengthy and constant visual monitoring of wave V latency is subject to human error. Medical errors may occur as a result of fatigue, poor communication, and inattention (Krueger, 2006). Automated monitoring algorithms may help avoid human error in intraoperative ABR monitoring, alerting the clinicians automatically if an abnormality in the ABR is detected. In the next section, automated diagnostic ABR analysis algorithms presented in the literature will be reviewed.

6.1.1.1.4 Analysing the Diagnostic ABR

As with the threshold ABR, analysis of the diagnostic ABR in clinical practice is based on the visual inspection of ABR waveforms by clinicians (British Society of Audiology, 2019b). This analysis is largely concerned with identifying the key components (waves) of the ABR waveform and measuring their latency (British Society of Audiology, 2019b). Increases in wave latencies or in the gaps between waves (inter-wave latencies) may indicate the presence of pathology (Hall, 2007). Wave labelling is not a straightforward procedure as even for normal ABR waveforms the morphology and latency of the peaks is heterogenous (Atcherson, 2012). In some cases the wave peaks may be clear and so the latency of the peak amplitude reflects the wave latency (Hall, 2007). However, there are several other instances when wave labelling is more challenging, e.g. in the case of fused/bifid/missing/extra or uncertain peaks (Hall, 2007) (Figure 6-2). Automated analysis of the diagnostic ABR may assist clinicians, in particular for challenging cases where the wave latencies are not clear. This will be particularly useful when assisting clinicians with less experience at interpreting diagnostic ABR data.

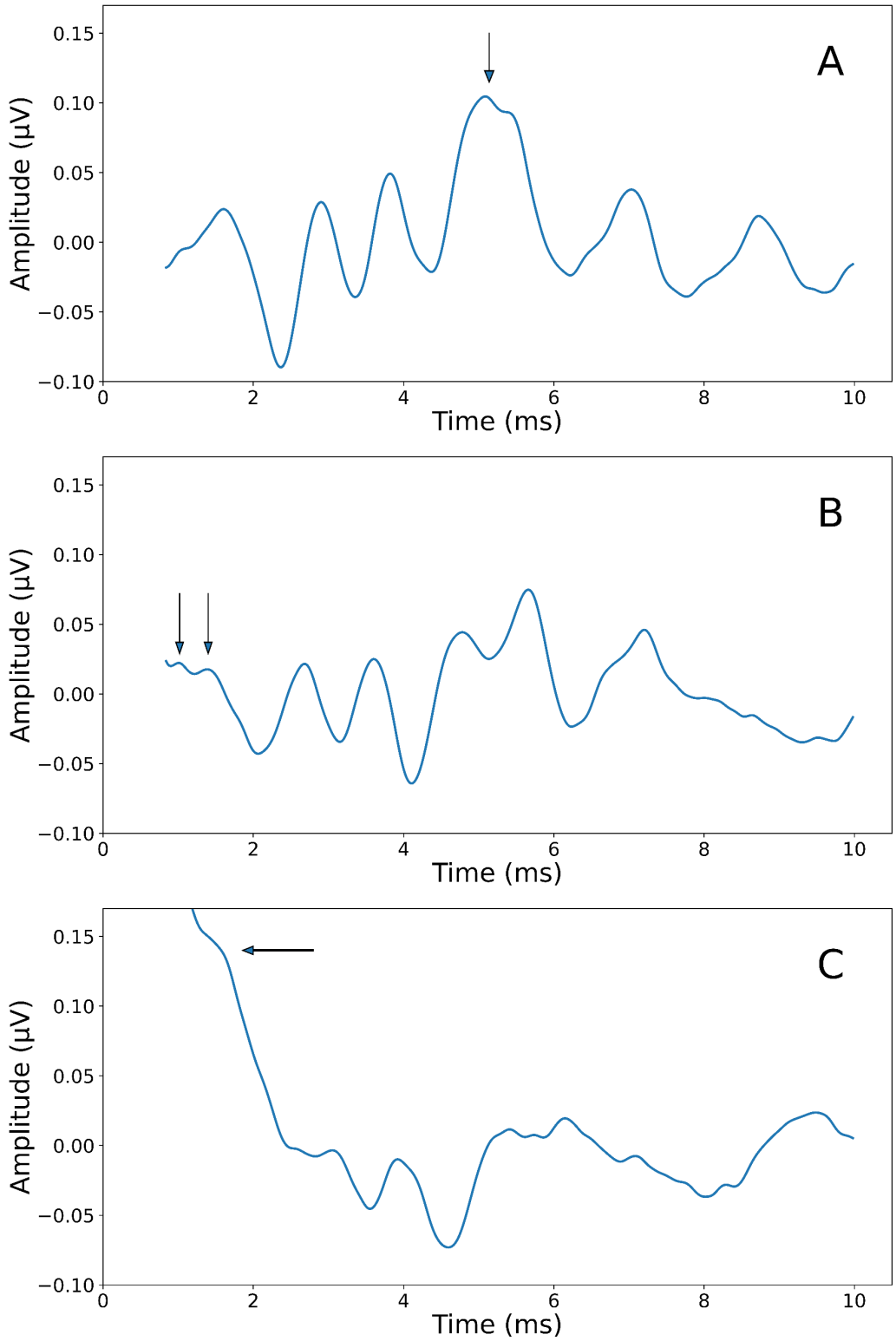

Figure 6-2 Examples of variable ABR morphology. Graph A shows a fused wave IV/V complex (arrow), with wave V appearing as a shoulder to the right of wave IV. Graph B shows a potential bifid (split in two) wave I (arrows). Graph C shows an ABR waveform where the morphology of wave I is unclear. These examples highlight the challenges faced by clinicians in analysing the diagnostic ABR.

6.1.1.1.5 The Potential Benefits of Automated Analysis

It has been pointed out that 'consistency in how peaks are marked in the clinic or laboratory is of significant importance' (Atcherson, 2012). This is because the ABR data must be analysed in the same manner as that used in defining the normative clinic data (Atcherson, 2012). Variability in the method used by clinicians to label the ABR data may result in waveforms falsely being classified as normal or abnormal. Automated ABR analysis algorithms may provide a method of standardising the analysis of the diagnostic ABR, providing clinicians with consistency, both when labelling a normative dataset and when labelling clinical data to be referenced with said normative data.

6.1.1.2 Automated Diagnostic ABR Analysis Algorithms

6.1.1.2.1 Rule-Based Algorithms

Prior to considering the machine learning approaches for analysing the ABR, it is useful to first consider the traditional 'rule-based' algorithms. These approaches may provide a performance benchmark by which to compare newly developed algorithms, as well as inform feature extraction techniques which may be used in conjunction with machine learning approaches.

In 1982, Fridman *et al.* used the zero-crossings of the first derivative of filtered waveforms to identify the peaks of the ABR. In a similar manner, Boston (1989) used zero-crossings of the first derivative to identify the ABR peaks. A rule-based algorithm was developed to label wave V based on parameters including peak-to-peak amplitude and peak latency (Boston, 1989). Wave V was correctly selected in 11/13 cases. The rule-based algorithm reported by Boston (1989), in addition to a wave V latency prediction, provided a confidence measure, rating the latency prediction as either being 'possible', 'probable', or 'certain'. Pool and Finitzo (1989) developed a rule-based algorithm to identify waves I, III and V. Performance was compared to that of two clinicians. A mean latency difference of 0.052 ms was reported between the proposed algorithm and labels provided by the two experts (it is not specified if this is the mean *absolute* difference or simply the mean difference). An example of the use of the first derivative to identify signal peaks is shown in Figure 6-3. Smoothing is first required to remove high-frequency noise content from the signal which may give rise to an inflated number of zero-crossings (Felinger, 1998). Parameters such as the amount of smoothing applied, minimum peak width, and minimum peak height may need to be set to avoid detection of false peaks (Felinger, 1998). The first derivative may be approximated using this (finite difference) equation (Felinger, 1998):

$$x'_i = \frac{x_{k+1} - x_k}{\Delta t} \tag{7.1}$$

where x_k is the k^{th} digitised sample of signal $\mathbf{x}$, and Δt is the difference in time between the two time points ($x_{k+1} - x_k$). A limitation of this method, as applied to ABR wave labelling, is that the ABR waves do not always correspond to voltage maxima (Atcherson, 2012). ABR wave labelling algorithms need to be able to account for missing, fused or bifid waves (Atcherson, 2012).

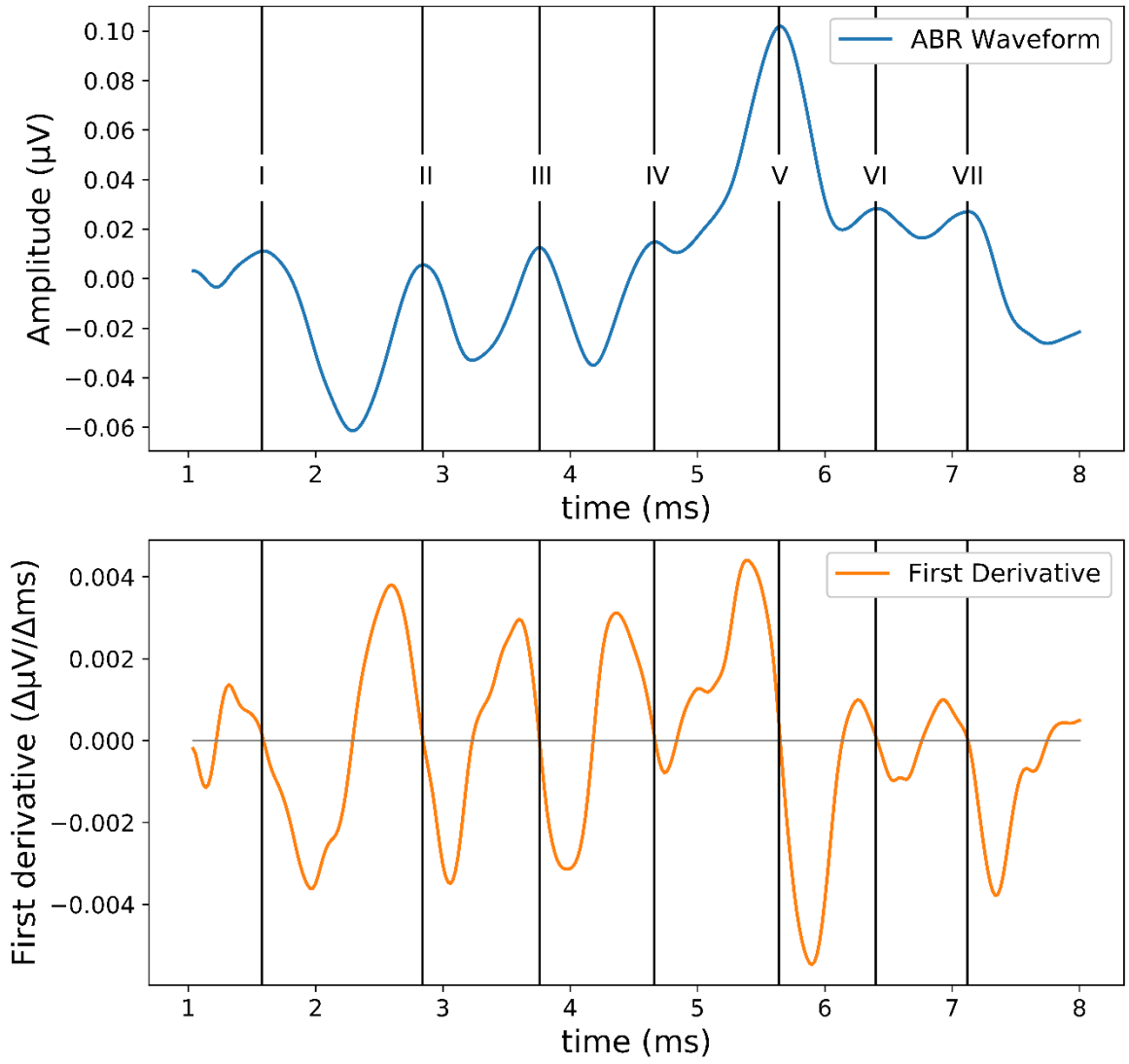

Figure 6-3 Use of the first derivative to identify signal peaks. The bottom graph shows the first derivative calculated from the ABR waveform presented in the top graph. The zero-crossings on a downward slope are marked with a vertical black line and correspond well to the latencies of waves I–VII of the ABR waveform. This approach works well in this example where the waves are nicely spaced and represent local voltage maxima. This is however not always the case.

Bradley and Wilson (2005) presented an algorithm to detect ABR peaks I to VII, using Gaussian wavelet analysis, combining signal smoothing and derivative estimation within the same operation (Bradley and Wilson, 2005). This was combined with a rule-based algorithm to estimate the ABR wave latencies. The algorithm performed well with mean absolute errors (MAEs) of 0.03,

0.05 and 0.06 ms for waves I, III and V respectively. A limitation of the proposed algorithm is that peaks were searched for in a set order and if a peak was not found then the search would stop, potentially leaving some peaks unidentified (Bradley and Wilson, 2005).

Other studies have also investigated the use of wavelet analysis to help identify the ABR wave latencies, e.g. work by Popescu *et al.* (1999) and Ikawa, Morimoto and Ashino (2014). Wavelet analysis is an effective biomedical signal analysis tool, helping to separate impulse-like events from diffuse EEG noise, and provides time-frequency localisation (Unser and Aldroubi, 1996).

Delgado and Özdamar (1994) presented an ABR peak identification and labelling algorithm which combined matched filtering with a rule-based approach. The first step was to apply a first order differentiating algorithm to the smoothed waveform in order to detect all of the waveform peaks and troughs. Time-shifts introduced by the matched filters were corrected for. Following matched filtering, which improves the SNR and enhances the ABR peaks (Delgado and Özdamar, 1994), a rule-based approach was implemented to label the ABR wave peaks (Figure 6-4).

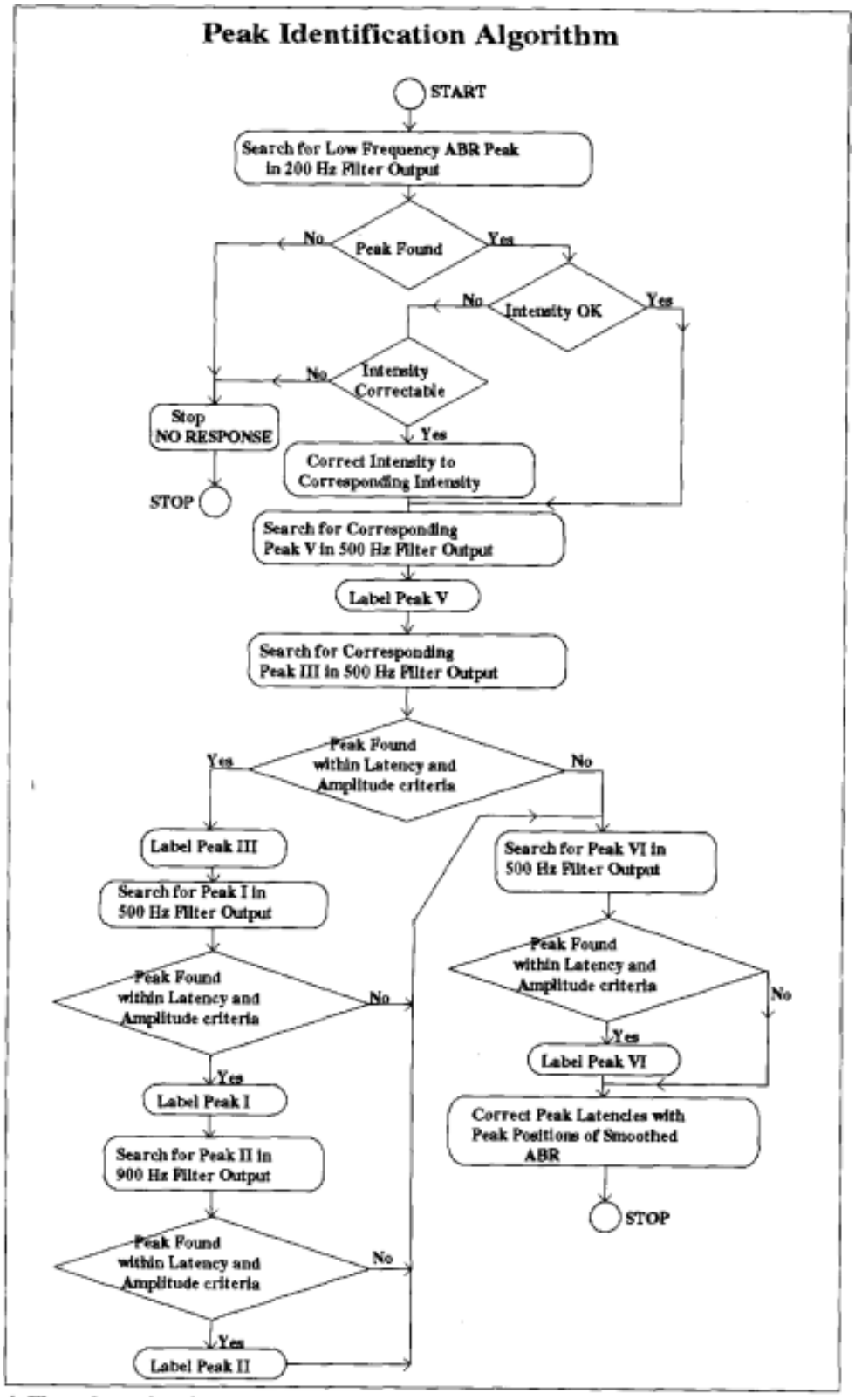

Figure 6-4 The rule-based ABR peak labelling algorithm presented by Delgado and Özdamar (1994). ABR waves I–V are labelled automatically using a combination of matched filtering and rule-based processing. Reproduced from Delgado and Özdamar (1994) with permission from IEEE (© 1994 IEEE). Note—a higher resolution image was not available.

As can be seen from Figure 6-4, the labelling of waves I–IV by the rule-based algorithm is done after wave V is labelled, allowing wave V to be used as a point of reference. This is similar to the approach that a human clinician may take, labelling the more obvious waves first and then using this information to help deduce the latency of the lower-amplitude waves. However, if wave V is incorrectly labelled, the effects of mislabelling will be cascaded down. Wave V was 'correctly labelled' for 96% of the data collected from individuals with normal hearing and 82.3% for data collected from individuals with a hearing loss. Two further methods in the literature making use of filtering to identify ABR peaks are provided by Pratt, Urbach and Bleic (1989) and Grönfors (1993). These references are provided for completeness, however, elaboration on the nature of these methods is omitted for the sake of conciseness.

Vannier *et al.* (2001) used a pattern matching algorithm to analyse the ABR, based on a technique developed by Motsch (1987). Using this method, a model of the ABR is generated. An ABR template is divided into four sections, with each section of the template being transformed in order to maximise the similarity (as measured by cross-correlation) between the ABR waveform and the transformed template (Vannier *et al.*, 2001). A related method using cross-correlation and template matching was previously presented by Elberling (1979). The latencies of waves I, II, III and V are then provided by the model where the 'true' latencies of the transformed template are known. Wave V was detected with an average deviation of 0.04 ms (SD 0.2 ms). Analysis of latency performance was not carried out for waves I, II and III.

Valderrama *et al.* (2014) used fitted parametric peaks to analyse the ABR. This method is based on the premise of template matching, However, a bank of templates to match with the signal is not required, as the template is synthesised for each individual signal analysed using parametric peaks generated to try to match the signal (Valderrama *et al.*, 2014). This approach was used to identify ABR waves III and V. These waves were searched for separately over a 3 ms window. This was used to ensure that adjacent waves weren't given the incorrect label whilst allowing leeway for latency shifts (Valderrama *et al.*, 2014). The disadvantage of this approach is that latency shifts outside of this window would not be detected. Unlike previous template matching studies which relied on subject recorded templates (Elberling, 1979; Motsch, 1987; Vannier *et al.*, 2001), this method synthetically derives templates under the assumption that ABR waveforms are 'shape-invariant' (Krumbholz, Hardy and de Boer, 2020). As discussed in Section 6.1.1.1.4, wave V morphology is highly variable, and the peak of the wave IV–V complex may not represent wave V. The assumption of 'shape-invariance' is therefore unlikely to be met (Krumbholz, Hardy and de Boer, 2020). Dynamic Time Warping (DTW) approaches do not rely on this assumption (Picton *et*

al., 1988 in Krumbholz, Hardy and de Boer, 2020). Krumbholz, Hardy and de Boer (2020) therefore opted to use a continuous version of DTW known as 'non-linear curve registration', overcoming the limitations of parametric template fitting techniques which use linear time shifts (Krumbholz, Hardy and de Boer, 2020). Continuous curve registration allows time warping of discretised continuous numerical data and ensures that each single point in the warped signal maps back to a single point of the original signal (Krumbholz, Hardy and de Boer, 2020). Bayesian analysis did not find any evidence to reject the null hypothesis of no difference being present between the manually and the automatically labelled wave latencies (Krumbholz, Hardy and de Boer, 2020).

6.1.1.2.2 Machine Learning Algorithms

In addition to traditional rule-based algorithms, several studies have investigated the use of machine learning algorithms to assist with automated ABR wave identification. Freeman (1992) identifies that rule-based algorithms require human experts in ABR labelling to be able to articulate the rules they use in order to translate them into an implementable algorithm, which can be challenging. Machine learning algorithms get around this by learning from the properties of the data how best to perform the task without any explicit need for defined *a priori* knowledge (Freeman, 1992). However, supervised machine learning algorithms do rely on labelled data from which to learn how to perform a task.

Freeman (1992) evaluated four different multilayer perceptron (MLP) architectures (each with three layers) for the purpose of identifying the latency of wave V. Separate training and test sets were used. In one of the experiments performed the derivative of the averaged waveform was used as a feature in addition to the averaged waveform itself. The task was evaluated as a binary classification task with the machine learning algorithm being considered to be correct if a positive prediction was made for a waveform where wave V was present, and the predicted latency was within 0.2 ms of the target label. The best-performing architecture achieved an accuracy of 85% (17/20 correct) on the test set data. The variation in test set performance between the five combinations of architecture/input features was minimal with accuracy ranging from 15–17 correct predictions out of 20. Due to the limited size of the dataset used it is difficult to have confidence in the generalisable performance of the proposed architecture. However, this study presents an early example of the potential promise that machine learning techniques have in interpreting the diagnostic ABR.

Habraken, van Gils and Cluitmans (1993) initially attempted to determine the wave V latency of ABR data using a single multilayer perceptron; however, performance was deemed to be poor using a test set of synthetic ABR data (mean test set scores of 33–63% wave V latency correctly identified, depending on the hyperparameters selected). An alternative approach was therefore

explored, making use of a series of small feature selection networks (single perceptrons or multilayer perceptrons) which incrementally selected smaller and smaller segments of the data, narrowing down the location of wave V before the final neural network made a final prediction of the wave V latency. Upon evaluation using subject recorded ABR data, this method achieved an 80% ± 6% SD agreement with the labels provided by a human expert.

Tian, Juhola and Grönfors (1997) used a four-layer MLP to label waves I–V. The averaged ABR data were filtered before applying principal component analysis (PCA), reducing the number of dimensions of the input data to the neural network from 472 to 15. The trained neural network was evaluated on a test set containing data from 37 ABR recordings. The reported mean error for wave V latency was -0.003 ms ± 0.093 ms (SD).

More recently, Chen *et al.* (2021) evaluated multiple deep neural network architectures in their ability to automatically identify waves I, III, and V of the ABR. An analysis window of 0–8 ms (321 discrete sample points) was used. The target labels used were a vector of the same length as the input data, marked as '1' where waves I, III, and V were deemed to be present, and otherwise marked as '0'. The four sample points before and after a labelled wave were marked as the 'characteristic area' of a wave. Seven machine learning architectures were evaluated using k-fold cross-validation ($k = 9$ folds) (Table 6-2).

Table 6-2 The neural network architectures evaluated by Chen *et al.* (2021), whereby the choice of a LSTM or bidirectional LSTM was evaluated as well as how many recurrent layers to use.

Neural Network Architecture
Single-layer LSTM.
Two-layer LSTM.
Single-layer bidirectional LSTM.
Two-layer bidirectional LSTM.
Three-layer bidirectional LSTM.
Four-layer bidirectional LSTM.
Five-layer bidirectional LSTM.

The neural network architecture incorporating three bidirectional LSTM layers was found to be most efficacious with a mean accuracy of 85.46% of wave label predictions being within ±0.1 ms, and 92.91% within ±0.2 ms, of the human-defined target labels. In addition to optimising the

network architecture, the authors sought to determine whether discrete wavelet pre-processing of the ABR data would boost performance further yet. Overall, and certainly for the best-performing neural network architectures, data pre-processing using the discrete wavelet transform was not found to be beneficial. The parameter value for the number of nodes in the hidden fully connected layer was also evaluated for the three-layer bidirectional LSTM (64, 128, 256, and 512 nodes). The best result (85.46% mean accuracy within ±0.1 ms of the target label) was obtained using the largest number of nodes evaluated (512). As the model architecture and number of hidden layer nodes were optimised based on their k-fold cross-validation performance, there is a potential that the best version of the model selected was overfitted to the data contained within these validation folds, i.e. the best model architecture was chosen based on its cross-validation performance and so the cross-validation score may no longer serve as a good representation of the model's generalisable performance. Using an additional set of test data or nested k-fold cross-validation (Varma and Simon, 2006) would help to overcome this limitation. Another limitation of the study is that other key hyperparameters do not appear to have been optimised. Such hyperparameters include the learning rate, batch size, dropout, activation function used, momentum, and the number of training epochs. It may well be that the performance of certain neural network architectures would be different had these hyperparameters been optimised. Having said that, it is unreasonable to expect every single hyperparameter combination to be evaluated, however, it would be beneficial to optimise values across a reasonable number of hyperparameters expected to most influence performance.

Based on the properties of LSTM networks and their demonstrated suitability in the analysis of biomedical time-series data (Ahmedt-Aristizabal *et al.*, 2018; Faust *et al.*, 2018), Chen *et al.* (2021) provide evidence to support the use of recurrent neural networks in automated ABR analysis. A gap in the literature exists in that other deep learning algorithms such as CNNs and CNN-LSTMs have not been explored for labelling the diagnostic ABR.

6.1.1.3 Summary of Algorithm Performance Presented in the Literature

Table 6-3 provides a summary of the various automated approaches presented in the literature (both rule-based and machine learning) for labelling the waves of the ABR.

Table 6-3 A summary of the various methods presented in the literature for automated ABR wave latency estimation is provided along with a summary of the reported results.

Study	Method	Results Summary
Boston (1989)	Rule-based algorithm using first order derivative. Provides a confidence output ('possible', 'probable', or 'certain').	Wave V 'correctly identified' (tolerance for a 'correct' prediction not specified) for 11/13 waveforms.
Pool and Finitzo (1989)	Rule-based algorithm.	Mean latency difference between algorithm and experts of 0.052 ms for wave V.
Delgado and Özdamar (1994)	Matched filtering combined with a rule-based algorithm.	Wave V 'correctly labelled' ('correct' definition not specified) for 96% of waveforms from individuals with normal hearing and 82.3% of waveforms from individuals with a hearing loss.
Vannier *et al.* (2001)	Pattern recognition/template matching.	Wave V detected with an average deviation of 0.04 ms (SD 0.2 ms).
Bradley and Wilson (2004)	Derivative estimation wavelets and a rule-based algorithm.	Mean absolute errors (MAEs) of 0.03, 0.05 and 0.06 ms for waves I, III and V respectively.
Kostorz *et al.* (2013)	IPAN99 rule-based algorithm which analyses angles and amplitude differences between sample points.	'Relative error' (definition not specified) of 1.04% for waves I, III and V.
Krumbholz, Hardy and de Boer (2020)	Non-linear curve registration.	Bayesian analysis found no evidence to reject the null hypothesis of no difference being present between the manually and the automatically labelled wave latencies.
Popescu *et al.* (1999)	Neural network-based filtering method to analyse wavelet	Wave V 'correctly' (definition not specified) identified in 92% of cases.

Study	Method	Results Summary
	transform maxima followed by rule-based localisation.	
Freeman (1992)	Multilayer perceptron.	A reported 85% accuracy (wave V correctly identified as being present/absent and, if present, the predicted label being within 0.2 ms of the target label).
Habraken, van Gils and Cluitmans (1993)	A series of perceptrons/multilayer perceptrons combined.	80% ± 6% SD agreement with the labels provided by a human expert for wave V latency, using subject recorded ABR data.
Chen *et al.* (2021)	A variety of recurrent neural network architectures were evaluated with a bidirectional LSTM being found to perform best.	85.46% of wave label predictions within ±0.1 ms, and 92.91% within ±0.2 ms, of the human-defined target labels.

6.1.2 Formulation of the Research Problem

There exist multiple advantages to being able to automate the process of labelling the diagnostic ABR. These include reducing interpretation time, reducing variability in interpretation, and providing equitable access to signal interpretation skills which is particularly helpful in settings where training or experience are limited. Additionally, automated detection may be of benefit during surgical monitoring where numerous sequential interpretations need to be performed over a long period of time (Chui, Murkin and Drosdowech, 2019).

Machine learning algorithms have been shown to be effective in a variety of biomedical signal processing tasks (Ahmedt-Aristizabal *et al.*, 2018; Hannun *et al.*, 2019) including in ABR detection (Alpsan, 1991; Acir, Özdamar and Güzeliş, 2006; Davey *et al.*, 2007; McCullagh *et al.*, 2007; McKearney *et al.*, 2022). Machine learning algorithms have the benefit of being able to learn how to perform the task from the data, potentially making use of features or properties of the data that may be challenging to identify and/or incorporate into a rule-based approach. Several studies have explored the use of machine learning algorithms to label the waves of the ABR (Freeman, 1992; Habraken, van Gils and Cluitmans, 1993; Tian, Juhola and Grönfors, 1997; Chen *et al.*, 2021). Several of these studies were published before the turn of the century and therefore do not make

use of the latest advances in machine learning, such as advances in convolutional neural network performance (Gu *et al.*, 2018), and were limited by the computational power available at the time. Methodological limitations also bring into question the generalisability of the findings of some of these studies. There is therefore a need for new research to explore and compare the performance of state-of-the-art machine learning algorithms using appropriate machine learning research methodology.

If the diagnostic ABR test is solely being used for a single clinical purpose, it may be sensible for the target label to be the presence of the pathology to be detected, e.g. is an acoustic neuroma present or absent. However, the diagnostic ABR may be used for a variety of clinical purposes, and it may be useful to provide automated wave labelling, with more detailed interpretation of the results left to the clinician who is aware of the wider clinical context of the test. This prevents pathologies, other than the primary target pathology, from being missed.

In terms of clinical usefulness, waves I, III, and V of the ABR are most important as these waves are the most robust (Atcherson, 2012). An automated detection algorithm should therefore prioritise and focus on being able to identify waves I, III, and V correctly. In addition to providing the wave latency values, it could be very useful for clinicians to be provided with a form of confidence measure, i.e. the degree of confidence that the machine learning algorithm has in being able to provide a reliable wave latency prediction. Providing a confidence measure for ABR latency predictions has received little discussion in the literature, with this review only identifying the work by Boston (1989) as having done so. An effective confidence measure would help clinicians better interpret the automated wave latency predictions and know how much weight to apply to these automated wave predictions when ultimately making the final decision using their overall clinical judgement.

6.1.3 Aims and Objectives

Aim 1. To propose, train, and evaluate automated machine learning algorithms which are able to label waves I, III and V of the diagnostic ABR. Multiple state-of-the-art algorithms should be evaluated to select the best approach. The automated algorithm should also provide a confidence measure to help clinicians interpret the latency values provided. The aim was not to present a final model, ready for clinical implementation, but rather to identify promising algorithms which may then be evaluated on larger datasets reflective of the intended clinical population.

Objective 1a: Prepare the dataset by labelling waves I, III and V of the data as well as providing a confidence value label for the latencies provided, using custom-built software.

Objective 1b: Compare machine learning algorithms using nested k-fold cross-validation.

6.2 Methods

6.2.1 Overview of Methods

Waves I, III, and V of diagnostic ABR waveforms from a previously recorded database were labelled by an audiologist. A confidence measure (range 0–5) was also provided. A selection of machine learning algorithms was trained and tested using a cross-validation approach, both in their ability to correctly predict the target wave latency and the confidence level.

6.2.2 ABR Data

The ABR data used in this secondary data analysis study were made publicly available by Sundaramoorthy *et al.* (2000). The link for accessing the data had become broken (http:\\www.engg.le.ac.uk\abrdata), however, the corresponding author of Sundaramoorthy *et al.* (2000) (Dr Michael Pont) was contacted and kindly provided permission to use the dataset and access to it. The dataset consists of suprathreshold ABR data recorded from 81 individuals (39 females; 42 males) aged 20–56 years with normal hearing (pure tone audiometry [PTA] thresholds ≤ 20 dB HL at octave intervals between 250–8,000 Hz, inclusive). ABR recordings were performed using a Nicolet Spirit Evoked Potential System in an electrically screened and acoustically isolated room (Stancold Acoustics). Recordings were made using silver chloride electrodes placed on the left and the right mastoid processes (A_1 and A_2, respectively), the forehead (common), and the vertex (C_z). Electrode impendences were <10 kΩ, with impedance differences between electrodes <5 kΩ. The participants were in a reclined position in a darkened room and in a relaxed/sleeping state for the recordings. An 80 dB HL 100-µs click stimulus was delivered to the test ear, with a 60 dB HL masking white noise delivered to the contralateral ear (Sundaramoorthy *et al.*, 2000). A stimulus rate of 10 Hz is reported in Sundaramoorthy *et al.* (2000); this recording parameter value is an unconventional choice given that 10 Hz is a subharmonic of UK mains alternating current (50 Hz). As the recordings appear free of mains artefact upon visual inspection, it is suspected that the reported nominal stimulus rate of 10 Hz is an approximate value. A 10 ms recording window was used, with a sampling rate of 50 kHz.

Further offline processing of these data in the current study included bandpass filtering between 100–3,000 Hz using a 3rd-order Butterworth filter. These filter settings reflect those recommended by the BSA (2019b).

For each of the 81 participants, both ipsilateral and contralateral recordings were made, using condensation, rarefaction, and alternating stimuli. Two repeat recordings of 1,024 recording epochs were made for each stimulus polarity/recording laterality combination, resulting in a total of 24 recordings for each of the 81 participants (Sundaramoorthy *et al.*, 2000). Only the ipsilateral recordings were used in the current study. The repeat recordings were averaged together to generate a grand average from 2,048 individual epochs, resulting in a final tally of six averaged ABR waveforms per participant. For one of the participants, four duplicate recordings appeared in the database. After removal of these duplicates, the database totalled 482 recordings.

6.2.3 Ethics

This secondary data analysis study was granted ethical approval by the University of Southampton Faculty Ethics Committee (ERGO 66305).

6.2.4 Data Labelling

Custom software was built using the Matplotlib Python library (Hunter, 2007) to allow the peaks of the ABR data to be labelled. One clinical audiologist (the present author) labelled waves I, III and V (Figure 6-5). Waves I, III and V were selected as these waves are the most important clinically, and also the most robust (Atcherson, 2012). Unlike other studies which have combined the inputs of multiple clinicians to label ABR waves (Chen *et al.*, 2021), the current study used the input of only one clinician. Using a group consensus labelling strategy has been shown to be effective, allowing machine learning models to classify data in a manner representative of the group of experts who labelled the data (Valizadegan, Nguyen and Hauskrecht, 2013). The machine learning approach presented by this current study is not intended to be used in its current state for clinical use, but rather to demonstrate the effectiveness a machine learning algorithm to be able to learn from the clinical acumen of an audiologist to label the waves of the ABR. Whilst using a single data labeller may be a potential limitation of the present study, using multiple labellers was considered superfluous for this early-stage study into the feasibility of using machine learning to label ABR waves. The presented machine learning approach may subsequently be trained on a larger dataset containing both normal and abnormal ABR data, labelled by a group of expert clinicians.

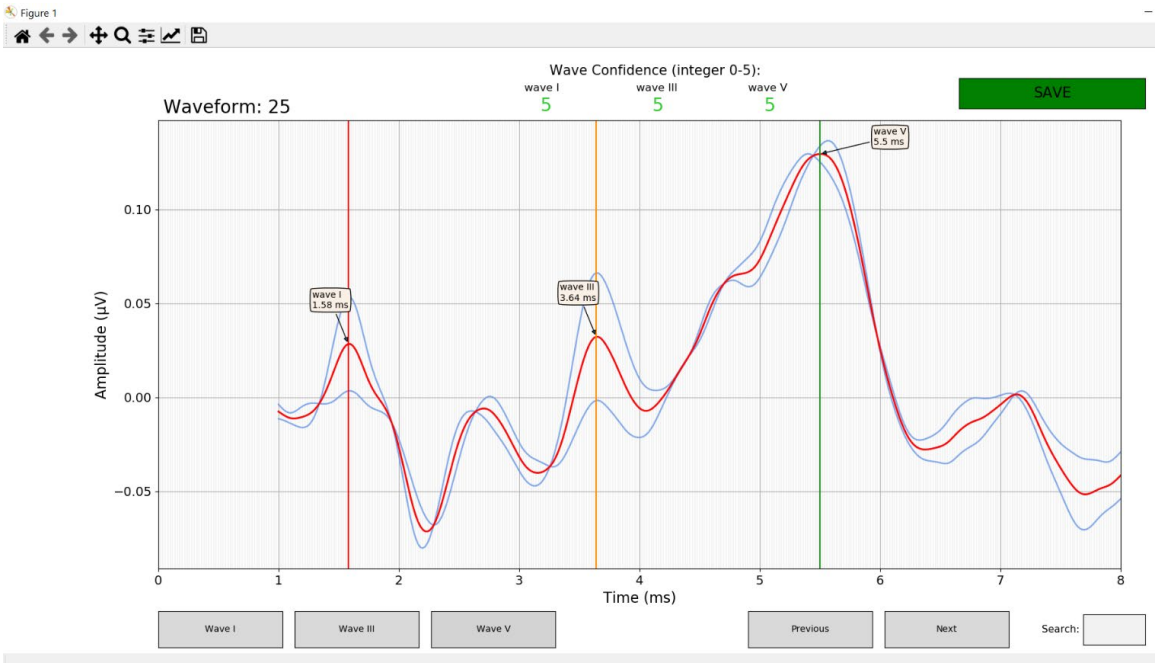

Figure 6-5 The custom software used to label the ABR waveforms. The custom labelling software presented each ABR waveform to be labelled (red), along with its two constituent sub-averages. This waveform represents one which the clinician had a high degree of confidence in labelling the latencies of waves I, III, and V, as the confidence label was five each of these waves. The zoom function allowed ABR waves to be labelled with a high degree of precision.

Using the custom-built user interface (UI), the audiologist was asked to label ABR waves I, III, and V, with the UI locking the latency label to the nearest sample point ($T_s = 0.02\ ms$). As well as labelling waves I, III, and V, the clinician was also asked to label the degree of confidence they had in their wave latency label, reflecting both their confidence in the wave being present, and also the labelled latency of said wave. A confidence rating was required for each of waves I, III, and V, and was provided as an integer between 0–5, inclusive. A description of the meaning of each of these ratings is provided in Table 6-4. Confidence labels may help clinicians interpret ABR waveforms, which are sometimes ambiguous. Having a confidence prediction provides the algorithm with the ability to deal with situations where an ABR is absent and alert the clinician to this possibility. This relatively novel feature of the methodology of the present study has not been used often in previous studies aiming to label the peaks of the ABR waveforms, apart from by Boston (1989). The confidence labels were additionally used to calculate sample weights when training the wave labelling algorithms, placing greater emphasis on training instances where the clinician had a high confidence in their wave latency labels (Byrd and Lipton, 2019).

Table 6-4 The confidence score descriptions used by the audiologist to label the confidence that they had in identifying and correctly labelling the latency of the ABR waves. Note how correct prediction of the latency is inherently linked to the ability to first correctly identify the presence of the wave in question.

Confidence score	Description
5	Very high confidence in being able to identify the wave correctly and estimate its latency.
4	High confidence in being able to identify the wave correctly and estimate its latency.
3	Reasonable confidence in being able to identify the wave correctly and estimate its latency.
2	Low confidence in being able to identify the wave correctly and estimate its latency.
1	Very low confidence in being able to identify the wave correctly and estimate its latency.
0	No confidence in being able to identify the wave correctly in order to estimate its latency.

The custom software for labelling ABR waveforms was designed in accordance with the principles of good user interface design (Tang and Patel, 1994; Patel and Kushniruk, 1998), in combination with the user-experience of an audiologist (the present author). Good clinical software UI should be simple for the user to utilise without having excessive features, feedback any errors to the user, and be structured in a way that is logical and intuitive to the user (Constantine and Lockwood, 1999). As both the designer and the intended user of the ABR labelling UI were one and the same, the software UI was readily able to be designed in a manner that met the needs of the user, through an iterative process of creation, UI testing, and improvement.

6.2.5 Machine Learning Approaches Evaluated

There is no single best machine learning approach for all tasks (Wolpert and Macready, 1997). One may therefore select potential machine learning approaches utilising *a priori* knowledge of the problem in combination with knowledge of machine learning approaches which may be suitable for the task, obtained from experience and from the literature. Given this information, several potential machine learning approaches with a track record in performing well for this kind of task were evaluated.

The machine learning algorithms were constructed using the Keras (Chollet and and others, 2015) Python software library.

6.2.5.1 Convolutional Neural Network

CNNs have been used extensively and to good effect for a variety of biomedical applications related to ABR wave labelling, including evaluating chromatographic peaks (Risum and Bro, 2019), electrocardiogram (ECG) QRS complex detection (Sarlija, Jurisic and Popovic, 2017), ABR detection (McKearney and MacKinnon, 2019; McKearney *et al.*, 2022), and EEG signal peak detection (Adam *et al.*, 2017). The architecture of the CNN is shown in Table 6-5.

Table 6-5 The convolutional neural network architecture. Optimised hyperparameters are shown in *italics* and underlined. Hyperparameters which were not fine-tuned are shown in regular typeface.

Layer	Hyperparameter settings
Separable Convolutional 1D	35 filters, *kernel size*, stride length=1, relu activation function, padding=same, kernel initialiser=he uniform
Max Pooling 1D	Pool size=2
Dropout	*Dropout rate*
Convolutional 1D	20 filters, *kernel size*, stride length=1, relu activation function, padding=same
Max Pooling 1D	Pool size=2
Dropout	*Dropout rate*
Flatten	

Layer	Hyperparameter settings
Dense	*Number of units*, *activation function*
Dropout	*Dropout rate*
Dense	120 units, *activation function*
Dense	3 units, linear activation

Hyperparameter optimisation was performed using a random search (using 15 hyperparameter combinations) of the hyperparameter space shown in Table 6-6.

Table 6-6 The hyperparameter space searched for the CNN.

Hyperparameter	Values searched
Dropout rate	0, 0.25, 0.5
Learning rate	0.00025, 0.0005, 0.00075, 0.001, 0.00125
Kernel size	5, 7, 9
Number of units	1500, 1750, 2000, 2250, 2500, 2750, 3000, 3250, 3500
Dense unit activation function	relu, selu
Number of training epochs	500, 800, 1100, 1400, 1700, 2000, 2300, 2600, 2900, 3200, 3500, 3800, 4100, 4400, 4700, 5000
Batch size	64, 128, 256
Loss function	Mean absolute error

6.2.5.2 Recurrent Neural Network / Bidirectional RNN

RNNs are particularly adapt to analysing sequential data such as biomedical signals due to the memory, or *state*, maintained from previously processed samples when considering the subsequent samples in the sequence (Chollet, 2018). The function of RNNs is considered in Section 4.2.8.2. There is a precedent set for the use of RNNs for biomedical signal peak detection: automated ECG wave labelling (Sampath and Sumithira, 2022), and EEG spike detection (Xu *et al.*,

2021). In the current study, LSTM units were used in the recurrent layers as these have been shown to be more adept at retaining information over longer time intervals than simple RNN units (Hochreiter and Schmidhuber, 1997).

RNN units will typically process a sequence unidirectionally from its beginning to its end, i.e. chronologically (Chollet, 2018). By evaluating a sequence in only one direction, RNNs may overlook patterns present in the data (Chollet, 2018). Bidirectional RNNs overcome this limitation by processing the sequence in both chronological and antichronological order (Schuster and Paliwal, 1997; Chollet, 2018). Chen *et al.* (2021) found bidirectional LSTM networks to be more effective than unidirectional LSTM networks at labelling the ABR waves, and so this approach was also evaluated in the current study. The architecture of the RNN used is shown in Table 6-7.

Table 6-7 The recurrent neural network architecture. Optimised hyperparameters are shown in *italics* and underlined. Hyperparameters which were not fine-tuned are shown in regular typeface.

Layer	Hyperparameter settings
LSTM **(x1, x2, or x3 layers)**	*Number of recurrent units*, *dropout rate*, *recurrent dropout rate* *(The number of recurrent layers was optimised in the random hyperparameter search to be either 1, 2, or 3 LSTM layers.)*
Flatten (only if returning sequences: optimised)	*Return sequences (True/False)*
Dense	*Number of units*, *activation function*
Dense	3 units, linear activation function

Hyperparameter optimisation was performed using a random search (using 15 hyperparameter combinations) of the hyperparameter space shown in Table 6-8.

Table 6-8 The hyperparameter space searched for the RNN.

Hyperparameter	Values searched
Number of LSTM layers	1, 2, 3

Hyperparameter	Values searched
Learning rate	0.00075, 0.001, 0.00125, 0.0015, 0.00175
Number of recurrent units	30, 40, 50, 60, 70, 80, 90, 100, 110, 120
Dropout rate	0, 0.1, 0.2, 0.3, 0.4
Recurrent dropout rate	0, 0.1, 0.2, 0.3, 0.4
Return sequences from last LSTM layer	True, False
Number of hidden Dense units	200, 300, 400, 500, 600, 700, 800, 900, 1000
Hidden Dense unit activation function	relu, selu
Number of training epochs	150, 200, 250, 300, 350
Batch size	128, 256
Loss function	Mean absolute error

The architecture and hyperparameter space searched for the bidirectional RNN algorithm were the same as those for the RNN (Table 6-7, Table 6-8), except bidirectional RNN layers were used instead of the standard (unidirectional) RNN layers.

6.2.5.3 CNN-LSTM Network

CNN-LSTMs combine convolutional layers with LSTM layers. This allows shorter sequences of features to be first extracted by the convolutional layer(s), before being further processed by the LSTM layer(s) (Chollet, 2018). LSTM algorithms have seen widespread use in biomedical signal analysis, including for foetal heart rate estimation from ECG (Fotiadou *et al.*, 2021), automatic detection of schizophrenia from EEG (Shoeibi *et al.*, 2021), and for seizure detection (Xu *et al.*, 2020). The architecture of the CNN-LSTM used is shown in Table 6-9.

Table 6-9 The convolutional long short-term memory network architecture. Optimised hyperparameters are shown in *italics* and underlined. Hyperparameters which were not fine-tuned are shown in regular typeface.

Layer	Hyperparameter settings
Separable Convolutional 1D	35 filters, *kernel size*, stride length=1, relu activation function, padding=same,
Max Pooling 1D	Pool size=2
Dropout	*Dropout rate*
Convolutional 1D	35 filters, *kernel size*, stride length=1, relu activation function, padding=same,
Max Pooling 1D	Pool size=2
Dropout	*Dropout rate*
LSTM	35 units, *dropout rate*, *recurrent dropout*, *return sequences (True/False)*
Flatten layer added if sequences are returned by LSTM layer	
Dense	*Number of dense units*, *activation function*
Dropout	*Dropout rate*
Dense	100 units, *activation function*
Dense	3 units, linear activation function

Hyperparameter optimisation was performed using a random search (using 15 hyperparameter combinations) of the hyperparameter space shown in Table 6-10.

Table 6-10 The hyperparameter space searched for the CNN-LSTM.

Hyperparameter	Values searched
Dropout rate	0, 0.1, 0.2, 0.3, 0.4, 0.5
Learning rate	0.0005, 0.00075, 0.001, 0.00125

Hyperparameter	Values searched
Kernel size	3, 5, 7
Number of Dense units	500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500
Dense unit activation function	relu, selu
Recurrent dropout rate	0, 0.1, 0.2, 0.3, 0.4
Return sequencers from LSTM layer	True, False
Number of training epochs	700, 900, 1100, 1300, 1500, 1700, 1900, 2100, 2300, 2500
Batch size	128, 256
Optimiser	Adam
Loss function	Mean absolute error

6.2.5.4 Multilayer Perceptron

MLPs consist of multiple layers of perceptrons and are considered in greater detail in Section 4.2.8.1. This type of machine learning algorithm has been previously used in the literature to label the waves of the ABR (Freeman, 1992; Habraken, van Gils and Cluitmans, 1993), and may potentially serve to act as a baseline for comparison. The MLP architecture used is shown in Table 6-11.

Table 6-11 The multilayer perceptron architecture. Optimised hyperparameters are shown in *italics* and underlined. Hyperparameters which were not fine-tuned are shown in regular typeface.

Layer	Hyperparameter settings
Flatten	
Dense	*Number of units in Dense layer 1*, *activation function*, L2 bias regularisation=0.01, L2 kernel regulariser=0.01
Dropout	*Dropout rate*

Layer	Hyperparameter settings
Dense	*Number of units in Dense layer 2*, *activation function*, L2 bias regularisation=0.01, L2 kernel regulariser=0.01
Dropout	*Dropout rate*
Dense	*Number of units in Dense layer 3*, *activation function*, L2 bias regularisation=0.01, L2 kernel regulariser=0.01
Dense	3 units, linear activation function

Hyperparameter optimisation was performed using a random search (using 15 hyperparameter combinations) of the hyperparameter space shown in Table 6-12.

Table 6-12 The hyperparameter space searched for the MLP.

Hyperparameter	Values searched
Dropout rate	0, 0.1, 0.2, 0.3, 0.4, 0.5
Learning rate	0.0008, 0.00085, 0.0009, 0.00095, 0.001, 0.00105, 0.0011
Number of units in Dense layer 1	150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 351
Number of units in Dense (hidden) layer 2	50 fewer units than that used in Dense layer 1
Number of units in Dense (hidden) layer 3	50 fewer units than that used in Dense layer 2
Dense unit activation function	relu, selu
Number of training epochs	Values between 500 and 2,500 inclusive in 50 epoch increments.
Batch size	16, 32, 64
Optimiser	Adam
Loss function	Mean absolute error

6.2.6 Input Features

Deep neural networks such as those outlined above are able to learn salient features from data in order to effectively perform a task. The coherent average was therefore one of the features provided as an input to these deep neural networks. Having said that, it can be useful to apply *a priori* knowledge of the task in order to extract useful features to be inputted to the neural networks. Feature extraction, can help reduce the time needed to train the algorithm and reduce the generalisation error by highlighting features most important to the task whilst ignoring/removing those which are noisy (Raschka and Mirjalili, 2017). Thus, the second and final input feature was the first derivative of the coherent average, which has proven in the literature to be a useful tool for ABR peak selection (Boston, 1989). An analysis window of 1–8 ms was applied to the coherent average and its first derivative, avoiding stimulus artefact and extraneous information. Each input feature was standardised separately within the cross-validation procedure described in Section 6.2.7. This involved subtracting the mean value from each feature vector and scaling to unit variance, using mean and standard deviation parameters derived from the training fold data.

6.2.7 Algorithm Evaluation using Nested K-Fold Cross-Validation

Nested k-fold cross-validation (Bergstra and Bengio, 2012) was used to evaluate the performance of the five chosen machine learning algorithms (CNN, LSTM, bidirectional LSTM, CNN-LSTM, and MLP) at predicting the wave latencies for waves I, III, and V. The task was treated as a regression task, with each algorithm having three outputs—each output predicting the wave latency of one of ABR waves I, III, or V. The predicted latencies were rounded to the nearest sample point. These machine learning algorithms were additionally compared to the baseline performance of a 'baseline regressor' (Pedregosa *et al.*, 2011), which ignores the input features and always predicts the mean latency values of waves I, III, and V from the training set data. A detailed description of nested cross-validation is described in Section 4.2.4, along with a diagram (Figure 4-7). For the current study, the outer loop of cross-validation was used to evaluate the generalisable performance of each of the machine learning algorithms. The inner loop was used to select the optimal hyperparameters using a random search of the available hyperparameter space. Due to computational limitations, only 15 separate hyperparameter combinations were explored for each machine learning model. The data were grouped into 81 groups, with each group comprising recordings made from one of the 81 participants. Each group was made up of six recordings apart from one group which had two (where the four duplicates were removed). Group k-fold cross-validation was used for both the outer and inner loops, with 27 folds in the outer loop and three

folds in the inner loop. There was therefore no participant overlap between the training and validation data portions.

6.2.8 Evaluation of the Best Algorithm for Confidence Label Prediction

Once the best-performing machine learning algorithm in terms of latency estimation was identified, the nested cross-validation procedure was repeated using the same algorithm, but this time using the wave confidence labels for waves I, III, and V instead of the wave latency labels. The nested cross-validation procedure was not repeated for all four machine learning algorithms due to the excessive computational cost that this would incur, and also on the basis that the best algorithm for ABR wave labelling would also have suitable properties for being able to link features of the ABR waveform to the confidence labels. Whilst a machine learning algorithm could have six outputs and have the dual task of predicting the three wave latencies and the three wave confidences, it should be noted that the confidence predictions would not relate specifically to the wave latency predictions made by the algorithm, but rather to the likelihood that an algorithm with a similar wave-labelling performance level to that of the human labeller in being able to label any specific wave with confidence, i.e. the ability of the algorithm to predict the same confidence levels as those labelled by the clinician for any given waveform.

Whilst the two tasks of wave latency estimation and confidence estimation are related, it was chosen to use a separate algorithm of the same type for each task, rather than one combined approach. One reason for this is that the optimal hyperparameter combination for each task may be different, and so combining the two tasks under the umbrella of a single algorithm with six outputs may lead to sub-optimal performance in one or both tasks. Another reason is that a single algorithm used for both tasks would likely require a greater number of parameters to be adjusted during training in order for it to learn six tasks instead of three. This could lead to the algorithm being harder to train and potentially being more prone to overfitting, reducing its generalisable performance. The confidence label predictions of the machine learning algorithm were rounded to the nearest integer to be congruent with the data labelling process.

6.2.9 Data Augmentation

Data augmentation was applied to the training folds within the cross-validation procedure (not the validation fold data). Data augmentation consisted of taking each training instance within the training fold and applying a scaling factor to increase/decrease its amplitude and also a latency shift, moving the waveform forwards or backwards in time along with the associated training wave latency labels. The amplitude scaling factors were randomly sampled from a normal

distribution with a mean of one and a standard deviation of 0.25. The degree of latency shift was randomly selected from a range of integers between -5 to 5, excluding zero, where the waveform was shifted either forwards or backwards in time by up to five sample points, i.e. ±0.1 ms. These parameter values were chosen to be sufficiently large so as to expose the machine learning algorithms to a more heterogenous set of training data, whilst not being so large as to make the data physiologically implausible, thereby making them unrepresentative of the other training data. Data augmentation doubled the amount of data in the training folds by producing one augmented waveform for each subject recorded waveform.

Note that data augmentation was not applied to the training data used by the 'baseline regressor', which simply predicted the mean latency values for waves I, III, and V, based on the subject recorded data in each training fold.

6.3 Results

In this section, the data from the ABR waveform labelling are reviewed first. Then the results of the wave latency estimation by the machine learning algorithms are reviewed. The confidence label estimation results will then be considered, before finally reviewing some example predictions, including examples where the machine learning algorithm performed poorly and where it performed well.

6.3.1 Data Labelling

6.3.1.1 ABR Wave Latency Labels

Figure 6-6 shows the distribution of the ABR wave labels as provided by an audiologist. It can be seen that the latency distributions for each of waves I, III, and V were quite narrow. Interpretation of the performance of the machine learning algorithms therefore needs to take into consideration the low variance of the latency labels.

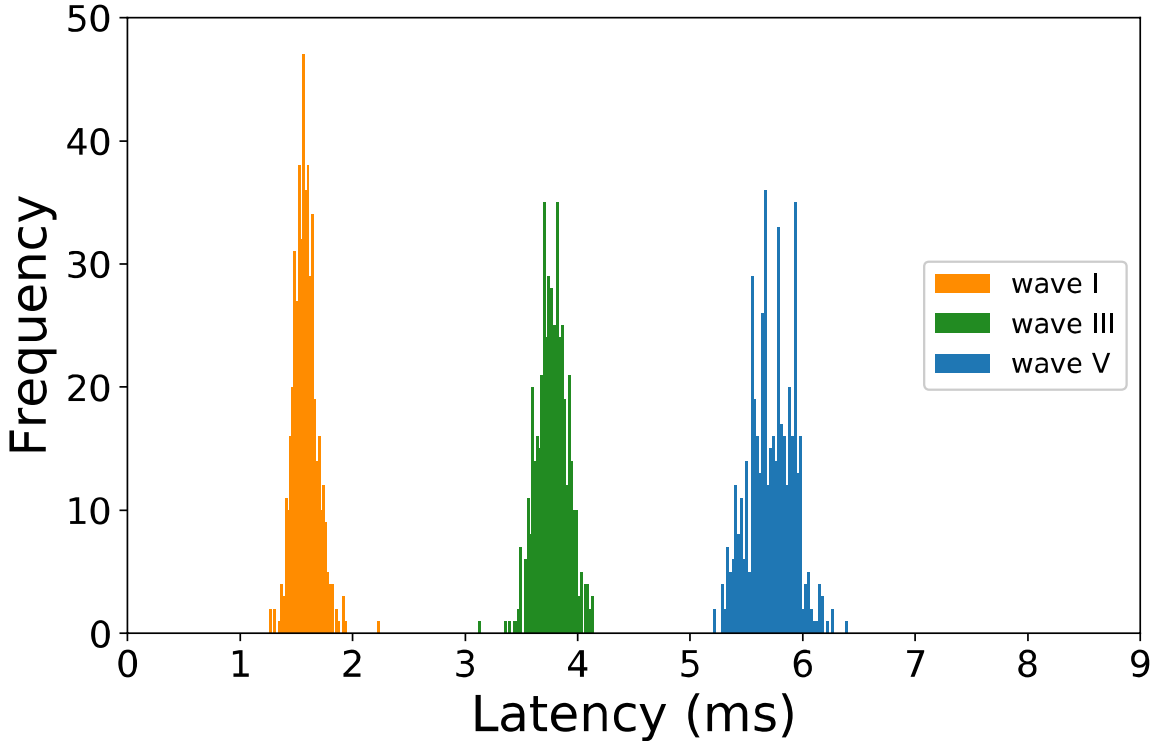

Figure 6-6 A histogram of the distribution of the ABR wave labels. The distributions of the ABR wave latencies as visually identified (the gold standard) were relatively narrow. The baseline regressor, which always predicted the mean latency values of waves I, III, and V as seen in the training data, provided a yardstick by which to compare the performance of the machine learning algorithms.

6.3.1.1.1 Confidence Labels

Figure 6-7 shows the distribution of the confidence labels for each of waves I, III, and V. It can be seen that for all three waves, the confidence level was generally high, with most confidence labels falling in the 3–5 range.

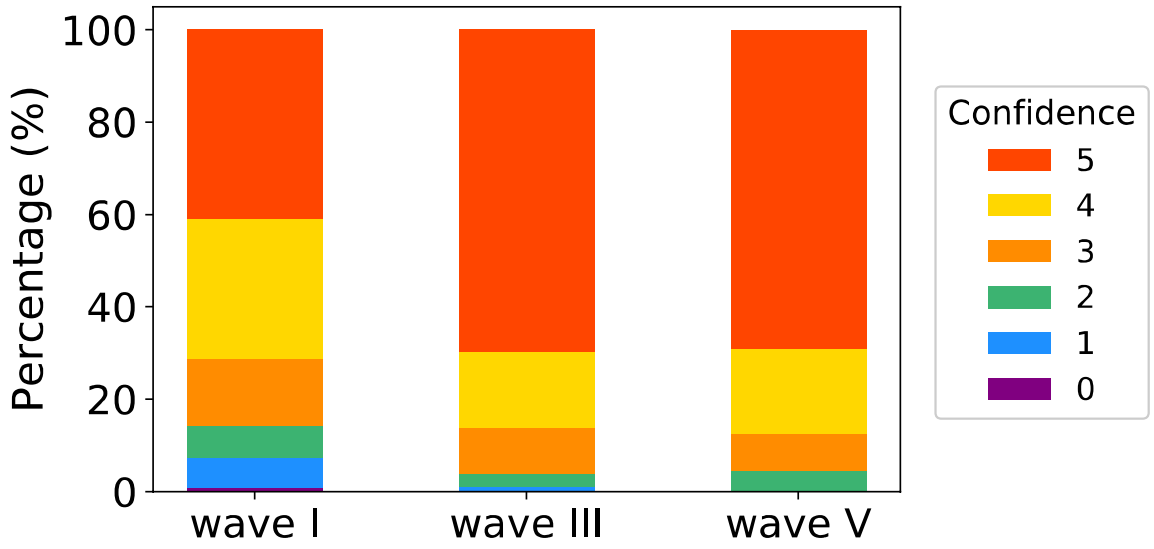

Figure 6-7 The confidence label distributions. These are broken down for each of waves I, III, and V.

6.3.2 Wave Latency Estimation

This section focuses on the wave latency estimation performance of the proposed algorithms. Nested k-fold cross-validation was used to evaluate the generalizable performance of the chosen machine learning approaches. Figure 6-8 shows the performance of the machine learning algorithms at predicting the latency of waves I, III, and V for data which had not been seen previously during model training. The performance of the machine learning algorithms was compared to that of the baseline regressor.

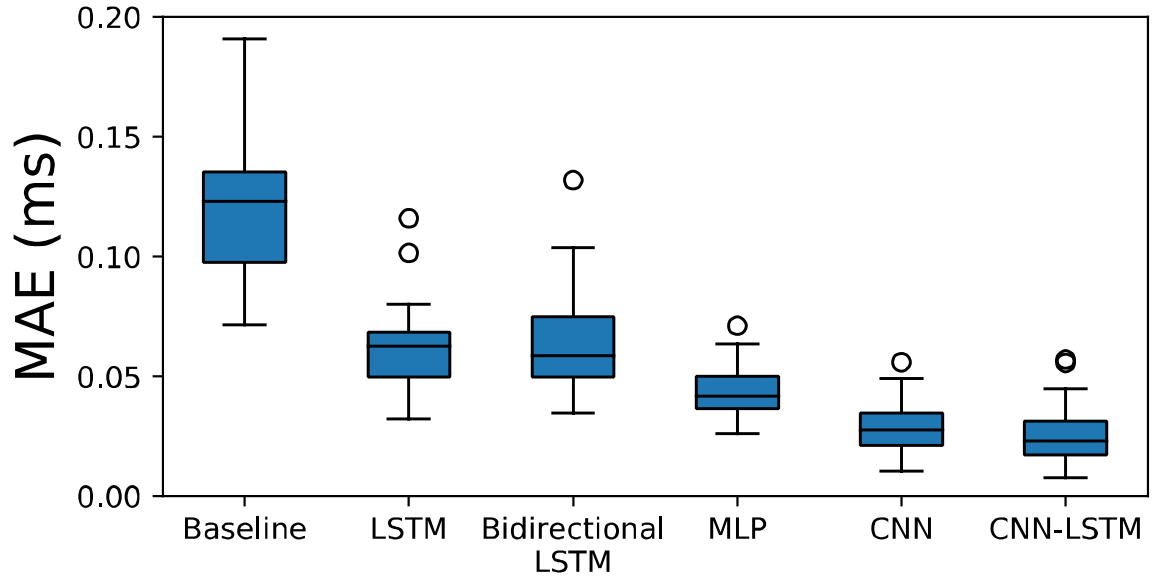

Figure 6-8 A comparison of machine learning algorithms for ABR wave latency estimation. These box-whisker plots show the mean absolute error (MAE) of the ABR wave latency

predictions for the outer validation fold data across the 27 outer loop iterations. The MAE scores include the combined performance across waves I, III, and V. The baseline is provided by the baseline regressor which simply predicted the mean latency value for each of waves I, III, and V, based on the training fold data. LSTM = long short-term-memory network; MLP = multilayer perceptron; CNN = convolutional neural network.

Figure 6-8 shows that even the performance of the baseline regressor was 'reasonable', with the baseline regressor achieving a median MAE of 0.123 ms across the 27 outer loop validation folds. This finding reflects the narrow variance of the ABR wave latency labels, which served to make the regression task easier to perform. Figure 6-6 shows the distribution of the latency labels for ABR waves I, III, and V in the dataset. This highlights the importance of providing a baseline performance measure, e.g. using a baseline regressor, which may help to make comparison between studies using different datasets easier. The best-performing algorithm was the CNN-LSTM, which had a median MAE of 0.023.

The distributions of the outer loop validation fold MAE scores were found not to be normally distributed for all of the algorithms, based on the visual inspection of probability plots and use of the Shapiro-Wilk test. The non-parametric Friedman's test was therefore used to test for differences between the outer loop validation fold scores between the algorithms. Post hoc testing was performed using the Wilcoxon signed-rank test, with correction for multiple comparisons made using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). A significant difference in latency prediction performance was found between algorithms: χ^2 (5) = 126, $p < 0.001$. The results of the post hoc testing are presented in Table 6-13. The best-performing algorithm was found to be the CNN-LSTM. It performed statistically significantly better than all other algorithms except the CNN (the next best-performing algorithm).

Table 6-13 The results of post hoc testing to compare latency estimation performance between the algorithms investigated. Correction for multiple comparisons was made using the Benjamini-Hochberg method. The table shows the corrected *p* values, with the significant findings in bold.

	Baseline	**LSTM**	**Bidirectional LSTM**	**MLP**	**CNN**	**CNN-LSTM**
Baseline						
LSTM	**< 0.001**					
Bidirectional LSTM	**< 0.001**	0.4004				
MLP	**< 0.001**	**< 0.001**	**< 0.001**			
CNN	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**		
CNN-LSTM	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	0.0585	

The CNN-LSTM achieved the lowest MAE for wave I and wave V, as well as the lowest overall MAE across all waves (Table 6-14). The CNN achieved the lowest MAE for wave III. As well as considering the MAE scores, it is useful to evaluate the percentage of latency predictions that occur within a given tolerance. This can help in evaluating the clinical utility of an algorithm which should both have a low mean error rate, and also a low number of outliers. Whilst a low mean error rate would likely indicate few outliers, this is not necessarily the case. Latency tolerance values of 0.1 and 0.2 ms were chosen as these have been previously reported in the literature (Chen *et al.*, 2021), allowing comparison between studies. Overall, the CNN-LSTM achieved the lowest average MAE score across outer validation folds, as well as the highest percentage of predicted wave latencies within 0.1 ms of the clinician-defined wave labels (Table 6-15). The CNN algorithm achieved the highest overall percentage of predicted wave latencies within 0.2 ms of the clinician-defined wave labels.

Table 6-14 A summary of algorithmic performance for the task of ABR wave latency estimation. The mean absolute error (MAE) scores provided are the average calculated across the 27 outer loop validation folds for each algorithm. Rather than calculating a macro average (arithmetic mean), the micro (weighted) average was calculated as one of the 27 folds contained two samples instead of six (weighted by the number of samples in each fold). The samples from the smaller fold were therefore weighted proportionally to the size of the fold for fairness. That being said, due to the large number of folds, the impact of weighting is minimal. The 'overall' column (light grey) represents the data for waves I, III, and V combined. The best score for each column is highlighted green for ease of comparison.

	Wave I MAE (ms)	**Wave III MAE (ms)**	**Wave V MAE (ms)**	**Overall MAE (ms)**
Baseline	0.084	0.111	0.164	0.120
LSTM	0.055	0.052	0.080	0.063
Bidirectional LSTM	0.055	0.061	0.081	0.066
MLP	0.046	0.035	0.051	0.044
CNN	0.030	0.016	0.039	0.028
CNN-LSTM	0.025	0.018	0.032	0.025

Table 6-15 The latency prediction performance for set tolerance levels. The scores provided are the micro-average calculated across the 27 outer loop validation folds for each algorithm. The 'overall' columns (light grey) represent the data for waves I, III, and V combined. The best score for each column is highlighted green for ease of comparison. Cochran's Q test showed a significant difference between the six algorithms evaluated, across each of the eight sub-columns in the table. Post hoc testing was performed using a pairwise McNemar test with correction for multiple corrections using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995), for each of the eight sub-columns in the table. ** This algorithm performed statistically significantly better than the other five algorithms for this given ABR wave and tolerance level. * There was no statistically significant difference between the performance of the multiple algorithms asterisked in this column, however, these asterisked algorithms all performed statistically significantly better than those not asterisked.

Percentage within a tolerance of:	**±0.1 ms**				**±0.2 ms**			
	Wave I	**Wave III**	**Wave V**	**Overall**	**Wave I**	**Wave III**	**Wave V**	**Overall**
Baseline	65.1	54.1	34.2	51.2	94.2	86.5	64.3	81.7
LSTM	85.3	85.5	72.4	81.1	97.3*	99.2*	94.2	96.9
Bidirectional LSTM	86.7	82.0	70.3	79.7	97.5*	98.5*	92.7	96.3
MLP	88.4	94.6	86.9	90.0	97.3*	98.3*	97.1*	97.6
CNN	92.3	98.3*	90.9	93.8	98.8*	99.6*	96.7*	98.3
CNN-LSTM	94.8**	98.5*	94.4**	95.9**	98.8*	98.8*	96.9*	98.1

6.3.2.1 Number of parameters

The hyperparameters and architecture of the machine learning algorithms were optimised within a nested cross-validation procedure (Section 6.2.7). As such the number of trainable parameters (weights) varied across the 27 outer folds. The mean number of parameters for each machine learning algorithm across the 27 outer loop validation folds are provided in (Table 6-16) as well as the standard deviation.

Table 6-16 The mean number of trainable parameters across all 27 outer loop validation folds for each of the machine learning algorithms evaluated.

Algorithm	Mean number or parameters (± SD)
LSTM	22,218,376 (4,312,155)
Bidirectional LSTM	40,863,369 (14,215,107)
MLP	253,294 (108,007)
CNN	6,121,346 (517,278)
CNN-LSTM	607,918 (765,450)

6.3.3 Confidence Level Estimation

As well as predicting the latency of waves I, III, and V, another aim of this study was to be able to predict the confidence that a clinician would have in being able to label a given ABR wave. As well as labelling the latency of all of the waveforms in the dataset, the audiologist (the present author) was tasked, for each of waves I, III, and V for each recording, with labelling the confidence level that they had in being able to identify the wave and label its latency accurately (Section 6.2.4). An integer rating was labelled in the range 0–5, inclusive. Nested k-fold cross-validation was performed again using the best machine learning algorithm at predicting the wave latencies (CNN-LSTM), however, the algorithm was this time trained instead using the confidence level labels to predict the confidence of a clinician (with the same ability as the clinician who labelled the data) in being able to identify and label an ABR wave for a given waveform. It should be noted that the confidence predictions of the machine learning algorithm do not relate specifically to the latency values predicted by the companion algorithm (separate CNN-LSTMs were trained for latency prediction and confidence prediction). Figure 6-9 shows how the predicted confidence levels of the machine learning algorithm related to the corresponding wave latencies predicted by the latency prediction CNN-LSTM for each waveform. The confidence levels predicted by the CNN-LSTM were rounded to the nearest integer.

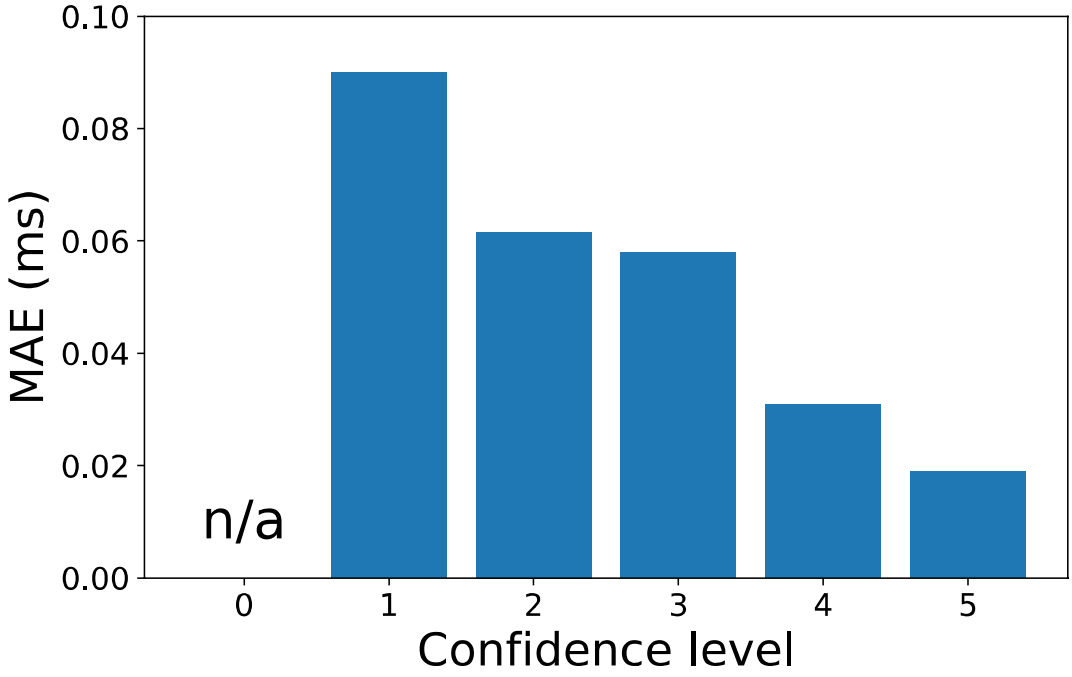

Figure 6-9 The average mean absolute error (MAE) is shown as a function of the predicted confidence level. It can be seen that as the confidence level of the machine learning algorithm increased, the error of the latency predictions decreased.

Figure 6-10 shows the percentage of wave latency predictions that occurred within a given tolerance (±0.1 ms and ±0.2 ms) of the target latency, for each level of confidence predicted by the CNN-LSTM. There were no confidence level predictions of 0 (n/a). There was a general trend that the higher the level of confidence predicted, the more accurate the latency predictions were, especially for a tolerance of ±0.1 ms.

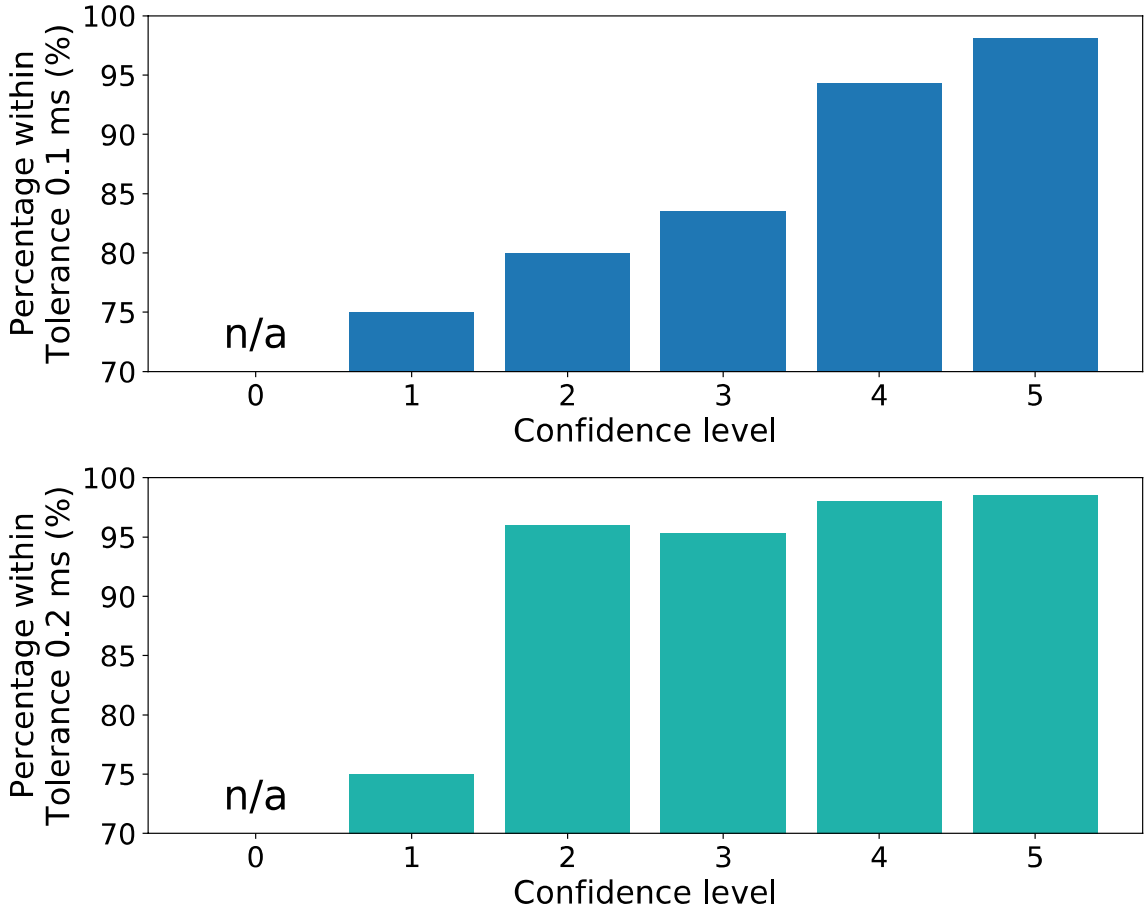

Figure 6-10 The percentage of latency predictions within a given tolerance of the target label are shown as a function of the predicted confidence level. The top graph (blue bars) shows the percentage within a tolerance of ±0.1 ms. The bottom graph (green bars) shows the percentage within a tolerance of ±0.2 ms. There were no confidence level predictions of 0 (n/a).

Overall, the predicted confidence levels corresponded reasonably well to the target confidence level labels provided by the audiologist. Spearman's rank correlation was calculated to measure the relationship between the predicted confidence levels and the target confidence level labels issued by the clinician. The predicted and target confidence levels were found to be positively correlated; $r(1444) = 0.53,\ p < 0.001$. Figure 6-11 provides a confusion matrix showing the relationship between the predicted confidence levels and the target confidence levels. It can be seen from Figure 6-11 that there were no predictions of 0 confidence, and that the majority of predictions and target confidence levels were in the 4–5 range.

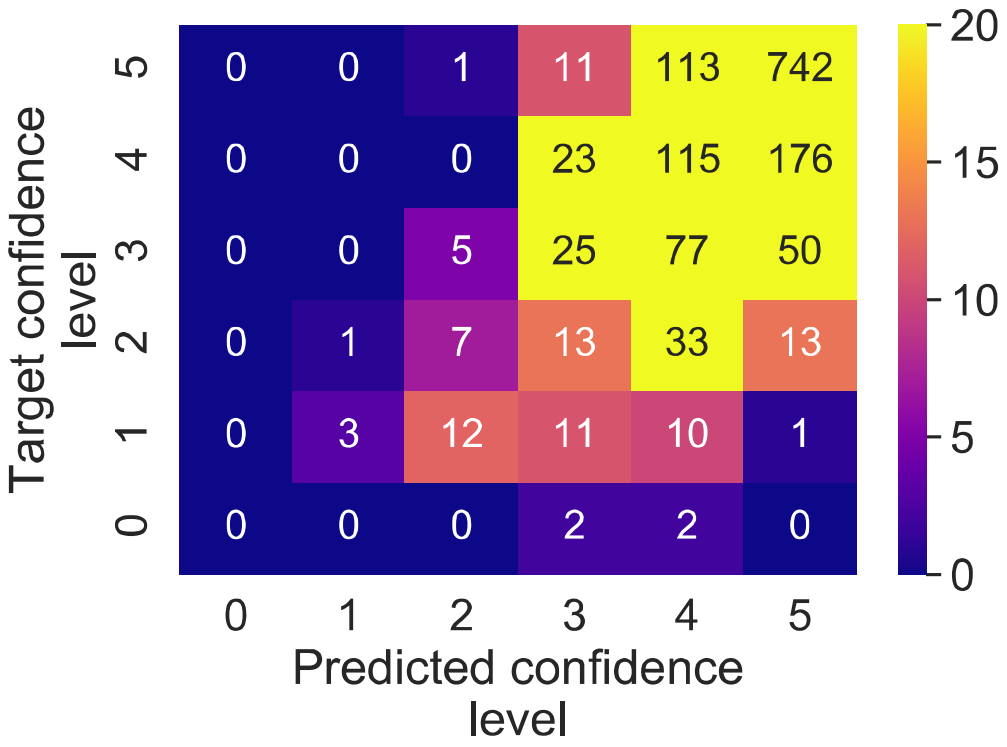

Figure 6-11 A confusion matrix showing the relationship between the predicted confidence levels and the target confidence level labels provided by the clinician.

6.3.4 Evaluation of Outlier Latency Predictions

In order to evaluate the proposed CNN-LSTM algorithms (one for wave latency prediction and one for confidence level prediction) in more depth, this next section will analyse those cases where the algorithm performed most poorly. The CNN-LSTM predicted one or more of waves I, III, or V as being outside of a tolerance of ±0.2 ms of the target label for 22/482 waveforms. Visual inspection was performed to determine the types of errors that the CNN-LSTM algorithm made. Eight broad categories of error were identified, including:

1. Wave V marked incorrectly as the peak of a wave IV/V complex where the right shoulder of the complex should have been marked.
2. Wave V incorrectly marked on the downslope after wave V instead of the peak.
3. Wave V error with wave V neither being marked as a peak or the shoulder of a wave IV/V complex.
4. The incorrect part of the shoulder of the wave IV/V complex was marked as wave V.
5. Wave I was bifid, with the incorrect location marked as the peak.
6. Wave I with downsloping morphology incorrectly marked.
7. Wave III error.
8. All waves (I, III, and V) marked to the left, i.e. earlier than the audiologist labels.

Error types 1–4 are shown in Figure 6-12.

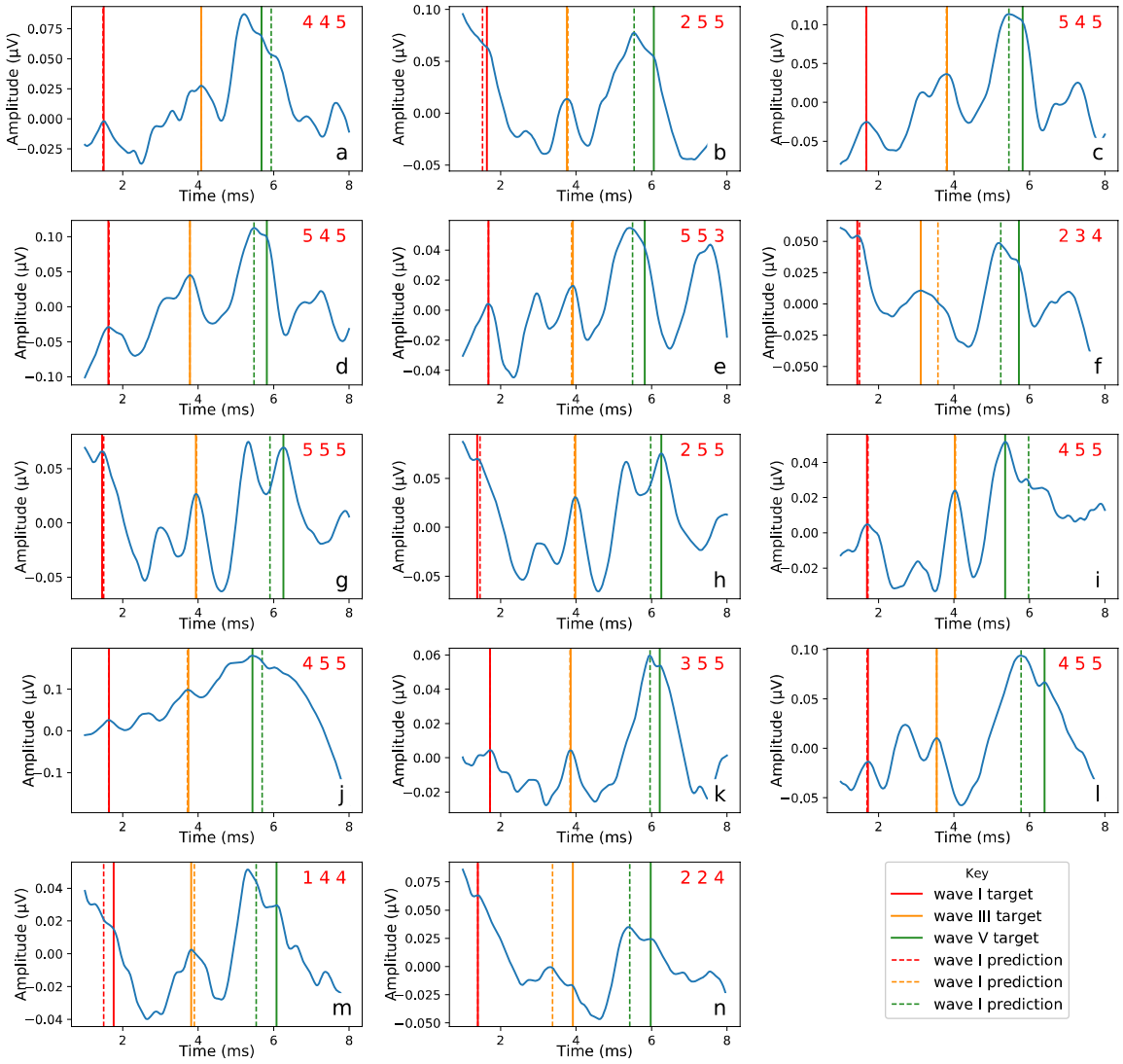

Figure 6-12 Outlier analysis—focussing on wave V errors. Plots (b), (c), (d), (e), (f), (k), (l), and (n) all depict type 1 errors, whereby wave V was incorrectly marked as the peak of a wave IV/V complex where the right shoulder of the complex should have been marked. Plots (f) and (n) also show type 7 errors, with wave III incorrectly marked. Plots (i) and (j) show type 2 errors, where wave V was incorrectly marked on the downslope after wave V instead of the peak. Plots (g), (h), and (m) show type 3 errors with wave V neither being marked correctly as the peak or the shoulder of a wave IV/V complex. Plot (m) additionally contains a type 5 error with a bifid wave I marked in the incorrect location. Plot (a) shows a type 4 error with the incorrect part of the wave IV/V complex shoulder marked as wave V. The confidence predictions for waves I, III, and V are shown in red in the top right-hand corner of each plot and suggest that, whilst there was an overall correlation between the predicted and target confidence labels, there were several examples where the algorithm predicated a high confidence level whilst being incorrect.

Error types 5–8 are shown in Figure 6-13.

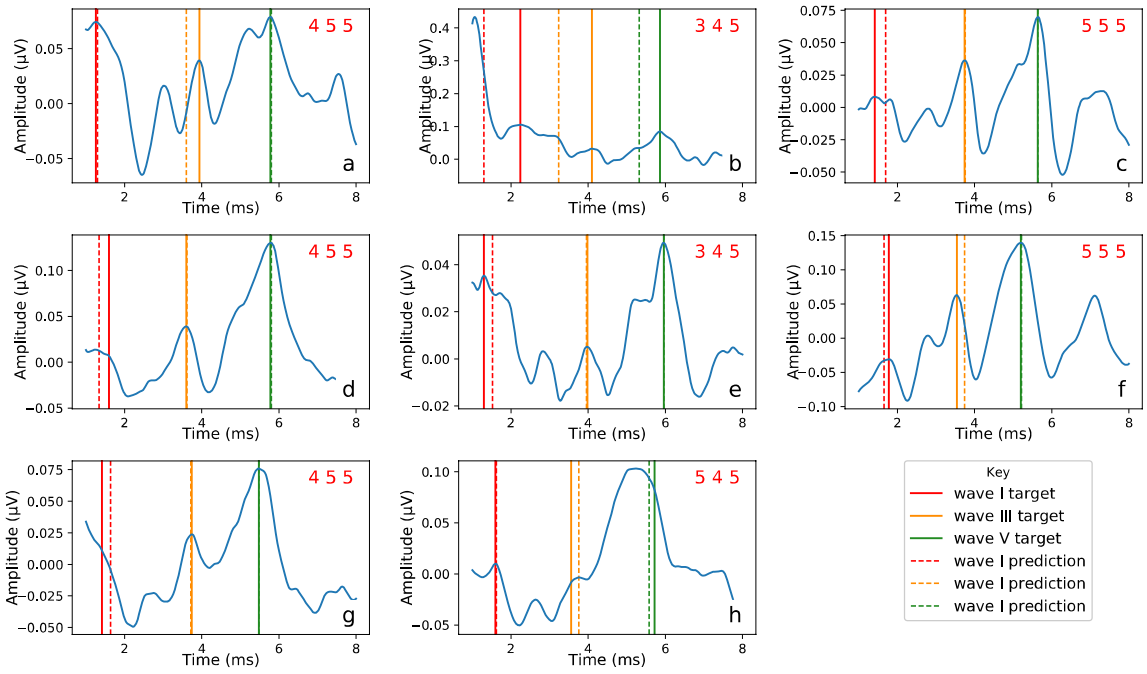

Figure 6-13 Outlier analysis. Plots (c), (d), and (e) show type 5 errors where wave I is bifid, and the incorrect location was marked by the neural network. Plot (g) shows a type 6 error where the incorrect part of wave I on a sloping baseline was marked. Plots (a), (f), and (h) show errors in the wave III latency prediction. Plot (b) shows a type 8 error where all wave predictions were shifted left of the target, i.e. earlier. The confidence predictions for waves I, III, and V are shown in red in the top right-hand corner of each plot.

6.3.5 Examples Where the Algorithm Worked Well

In order to provide some balance, this section will provide some examples of where the latency and confidence prediction algorithms were effective. Whilst Section 6.3.4 focuses on errors made by the algorithm, it should be noted that for the best-performing algorithm (CNN-LSTM), 95.9% of latency predictions were within 0.1 ms of the clinician-defined label. Figure 6-14 provides an illustration of some examples where the machine learning algorithm performed well.

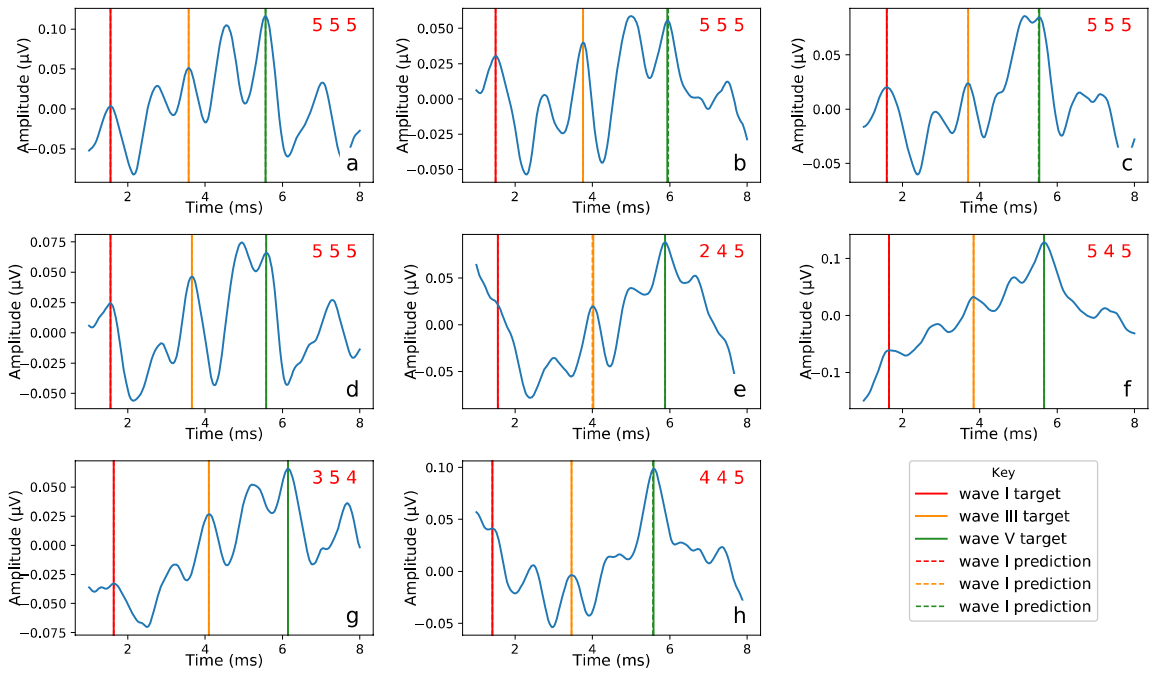

Figure 6-14 Examples where the CNN-LSTM performed well. In these examples, the error in each wave latency prediction is ≤0.02 ms. Some of these examples include tricky cases: plot (c) shows a wave IV/V complex where wave V is lower in amplitude than wave IV, plot (e) shows a wave I on a sloping baseline, and plot (g) shows a bifid wave I.

6.4 Discussion

6.4.1 Wave Latency Estimation

The primary aim of the current study was to propose, train, and evaluate automated machine learning algorithms which are able to label waves I, III and V of the diagnostic ABR. Of the machine learning algorithms evaluated, the best machine learning algorithm (a CNN-LSTM) was able to label previously unseen data, on average, with a MAE of 0.025 ms, with 95.9% of latency predictions being within 0.1 ms of the target. A baseline regressor, which simply predicted the mean values of waves I, III, and V from the training data, was used to provide a baseline performance level by which to compare the performance of the machine learning algorithms. The average MAE of the baseline regressor was 0.12 ms—almost five times larger than the MAE achieved by the CNN-LSTM. The performance of related algorithms reported in the literature may also provide a benchmark level by which to compare the performance of newly proposed algorithms. That being said, there are multiple factors which make drawing meaningful comparisons between studies challenging. Namely, datasets may be heterogeneous, recorded from different populations (e.g. including both normal and abnormal data) with different variance of the target variable (wave I, III, and V latency) within the data. Indeed, studies may even use

different target variables, labelling different waves of the ABR, e.g. wave V only, or including also waves II, and IV in the analysis. Additionally, the outcome measures reported by studies may be vastly different. These factors can make it extremely challenging to draw comparisons between studies. There are several steps that studies may take to help reduce this variability and allow comparisons to be drawn more readily. Such steps include recording and processing the data in a standardised manner in accordance with the prevailing national/international guidelines; using standardised target variables (e.g. waves I, III, and V) as used commonly in clinical practice and reported in the majority of studies (additional target variables may be included on top of the standard ones, e.g. waves II, and IV); using standardised outcome measures to report results, or even a range of outcome measures; reporting the variability of the target variable within the dataset, or reporting the performance of a baseline estimator; and making the dataset available (where possible) to allow subsequently developed algorithms to be evaluated and compared to previous algorithms.

Previous studies using machine learning methods to label the waves of the ABR include Freeman (1992); Habraken, van Gils and Cluitmans (1993); and Chen *et al.* (2021), (Table 6-3). Freeman (1992) used a multilayer perceptron and achieved an accuracy of 85%, where wave V was correctly predicted as present/absent, and if present, the predicted latency was within 0.2 ms of the target latency. In the current study, the overall best algorithm (CNN-LSTM), was able to correctly predict the latency of wave V within 0.2 ms for 98.1% of waveforms. This represents a significant improvement in performance over that previously reported, possibly due to the advances in machine learning techniques over the past decades, although acknowledging the caveats made previously for why drawing comparisons between studies is challenging. Notably, Chen *et al.* (2021) use a dataset which contains both normal and abnormal ABR data whereas the present study uses only ABR data recorded from individuals with normal hearing. Habraken, van Gils and Cluitmans (1993) used a series of perceptrons/multilayer perceptrons combined and achieved an 80% ± 6% SD agreement with the target labels provided by a human expert for wave V latency. The tolerance for a correct prediction is not provided, and so comparison of performance with that presently reported is not readily possible. One of the most recent and best-performing algorithms reported (including both machine learning and traditional rule-based algorithms) is that by Chen *et al.* (2021). They reported the performance of various configurations of recurrent neural networks using either unidirectional or bidirectional LSTM layers. The best performance reported by Chen *et al.* (2021) was an accuracy of 85.5% and 92.9% of wave latency predictions within 0.1 ms and 0.2 ms of the target label, respectively. This level of performance was achieved using a neural network with three bidirectional LSTM layers. In the current study, the CNN-LSTM achieved an accuracy of 95.9% and 98.1% of wave latency predictions within

0.1 ms and 0.2 ms of the target label, respectively. This represents an improvement upon the performance level reported by Chen *et al.* (2021). Again, it is important to acknowledge the heterogeneity between datasets. Unlike the dataset used in the present study, the dataset used by Chen *et al.* (2021) included data from individuals with hearing loss (as well as from individuals with normal hearing). This needn't mean that the ABR morphology would be abnormal but would likely serve to increase the variability of latency values for the ABR waves in the dataset. This is because the latencies of the ABR waves are affected by the degree of cochlear hearing loss (Jerger and Johnson, 1988). This being said, the CNN-LSTM performed better than the bidirectional LSTM on the same dataset in the current study, suggesting that the use of convolutional layers may yet further improve algorithm performance. However, this performance may have been influenced by the differences in the number of trainable parameters between algorithms, with the CNN-LSTM having far fewer trainable parameters than the bidirectional LSTM. Note that the number of trainable parameters was optimised within the nested cross-validation procedure and so varied across machine learning algorithms and even across outer loops cross-validation folds for the same algorithm. Bidirectional LSTMs process sequences in both chronological and antichronological order allowing additional representations of the data to be learned (Chollet, 2018). Whilst this may confer some performance benefit, the number of parameters in the layer are doubled (Table 6-16), potentially making training the network more challenging and leading to overfitting (Chollet, 2018). There was no statistically significant difference in performance between the bidirectional LSTM and the unidirectional LSTM in the present study (Table 6-13), despite a large difference in the number of trainable parameters (Table 6-16). There was also no statistically significant performance between the CNN and the CNN-LSTM when considering the latency estimation performance (Table 6-13). However, the CNN had far fewer trainable parameters on average than the CNN-LSTM. It could be that the CNN could perform better if it had more trainable parameters available, however, the number of parameters were optimised within the confines of nested k-fold cross-validation in terms of the kernel size and the number of units in the first dense hidden layer (Table 6-6) and these optimised hyperparameters were not always the largest values of the hyperparameter space explored.

6.4.2 Confidence Labels

Part of the aim of this study was to be able to provide a confidence measure to help clinicians interpret the ABR wave latency predictions provided. This feature is quite novel with only the study by Boston (1989) identified in the literature review as providing a confidence measure for wave latency predictions. Understanding the degree of uncertainty associated with a prediction can help clinicians know how much emphasis to allocate to it and to know when the prediction

should likely be discounted or at least interpreted with considerable caution (in the event of low confidence). Providing confidence predictions is a challenging aim to fulfil as the level of certainty regarding a latency prediction is challenging to quantify. The method adopted in the current study to achieve this aim was to train a second LSTM-CNN (chosen as this was the best identified algorithm for wave latency estimation) and to train it using the confidence labels provided by the clinician. This means that the predicted confidence level does not relate specifically to the latency predictions provided by the latency prediction algorithm, but rather to the level of confidence that one may place in a regression algorithm, of a similar ability to the clinician who labelled the ABR data, to be able to accurately predict the correct wave latencies for any given waveform. In other words, the confidence level prediction algorithm is trying to predict the level of confidence that the human clinician (who labelled the data) would have when labelling the latency of the ABR wave in question, which is expected to correlate to wave latency prediction accuracy. An alternative approach to providing a measure of confidence in the wave latency predictions could be to use deep Bayesian neural networks or Gaussian processes (Li *et al.*, 2021). Li *et al.* (2021) state that 'point predictions in absence of uncertainty estimates lack credibility quantification and raise concerns about safety'. Understanding and dealing with uncertainty is an important part of medical care. ABR detection in clinical practice uses confidence measures such as the Fsp or the Fmp, from which an associated *p* value may be obtained to aid the clinician's interpretation of the data. Such a measure would be useful also for wave latency predictions. Whilst the predicted and target confidence levels were found to be positively correlated ($r(1444) = 0.53,\ p < 0.001$), the majority of the wave latency labels fell into the higher confidence range (4–5 out of 5). This means that the algorithm may not have been exposed sufficiently to data where low clinician confidence was present and therefore may not be able to recognise such cases effectively.

6.4.3 Evaluation of Outlier Latency Predictions

Figure 6-12 and Figure 6-13 display the ABR waveforms where one or more of the predictions for the wave I, III, and V latencies were outside a tolerance of ±0.2 ms of the target label provided by the clinician. The morphology of the ABR varies between individuals, making the task of labelling its waves challenging. Machine learning algorithms trained on too small a dataset may therefore not be exposed to the wide variety of ABR morphologies present in the general population. The key themes in the error analysis were the machine learning algorithm mistakenly labelling the peak of a wave IV/V complex instead of the shoulder (wave V), and also selecting the incorrect peak for wave I when it was bifid. These types of morphology may be challenging for clinicians to label correctly and so it is understandable, especially given the size of the dataset, how the machine learning algorithm may also struggle with these types of ABR morphology.

6.4.4 Limitations and Future Work

A major limitation of the current study is that the dataset was obtained from individuals with no auditory pathology, i.e. all ABRs had a normal morphology, albeit in the presence of inter-subject waveform morphology variability. All of the participants had normal hearing. Whilst this study serves to compare the effectiveness of various machine learning approaches in labelling the ABR, an algorithm designed for implementation in the clinical setting would need to be trained and tested on a heterogeneous dataset which contained both normal and abnormal data, reflecting the characteristics of the target clinical population. The current study identifies suitable algorithms, accurate at labelling the normal ABR data, which may be suitable for use in future work.

A further limitation of the current study is that only one clinician was used to label the ABR data, as opposed to a group of clinicians as was the case in the study by Chen *et al.* (2021). The impact of this limitation is relatively low in the present study, as the main aim related to how well a machine learning algorithm was able to mimic the wave labelling process of one audiologist, rather than a consensus of audiologists. If the algorithm were to be used in clinical practice it is essential that the data are labelled by a group of experts with significant experience in ABR interpretation so that the algorithm may be able to learn to mimic this 'gold standard' level of performance.

Another limitation is that no previous traditional rule-based algorithms were implemented for comparison with the machine learning methods evaluated. The main reason why this was not performed is that, despite the publication of methods, there is often insufficient detail available to reproduce an algorithm in its entirety so that it is implementable in exactly the same format as that used in the published study. Additionally, certain algorithms may be set up for certain recording parameters and require adaptation for use on new datasets. One method of overcoming this limitation is to make datasets openly available for future researchers to use. A baseline regressor was used to provide a performance benchmark.

6.5 Conclusions

The main aim of this study was to train a machine learning algorithm to predict waves I, III and V of the diagnostic ABR. Of the algorithms compared, the CNN-LSTM performed the best and was able to label 95.9% of ABR waves within 0.1 ms of the target label. The average MAE was 0.025 ms. This exceeds the state-of-the-art level performance reported in the literature; however, it is acknowledged that meaningful comparisons between studies can be difficult to make due to heterogeneity between the datasets used. This study also reports a novel method of estimating

the uncertainty associated with ABR wave latency predictions. Further research using a larger heterogenous dataset including abnormal ABR waveforms is required before the presented machine learning algorithm may be implemented for use in the clinical setting. This is important in order to evaluate whether the proposed algorithm will perform well across the range of data likely to be encountered in the clinical setting. A carefully trained algorithm has the potential to be able to assist clinicians in the complex task of analysing the diagnostic ABR.

Chapter 7 Conclusions

The overarching theme of this PhD thesis is improving objective analysis of the auditory brainstem response. This has been achieved through work focussing on three key areas where gaps in the literature were identified:

1. Improving ABR detection using machine learning (Chapter 4).
2. Optimising weighted averaging and combining this with statistical ABR detection (Chapter 5).
3. Improving automated analysis of the diagnostic ABR using machine learning (Chapter 6).

7.1 ABR Detection using Machine Learning

Existing studies evaluating the use of machine learning to detect the ABR have demonstrated promising results. Improving ABR detection has the potential to benefit national newborn hearing screening programmes and also assist clinicians whose interpretations presently are reliant upon the visual inspection of often ambiguous waveforms. A significant limitation of studies in the field of ABR detection using machine learning is that the datasets used are often relatively small compared to those datasets used in the wider field of machine learning, e.g. 285 recordings (Alpsan *et al.*, 1994), 550 recordings (Davey *et al.*, 2007), 488 recordings (Acir, Erkan and Bahtiyar, 2013), and 810 recordings of 64 epochs each (R Zhang *et al.*, 2005). Additionally, it is very challenging to make meaningful comparisons between studies due to the varied datasets and outcome measures used. The research presented in this thesis aimed to overcome these limitations by using simulation to generate a large dataset with known labels by which to effectively train and evaluate machine learning algorithms. Having compared the performance of a range of machine learning algorithms, the proposed stacked ensemble was presented as an effective algorithm for ABR detection. Prior to this work, it was not known how the performance of machine learning algorithms compared to that of statistical ABR detection methods such as those used in commercially available evoked potential equipment.

The hypothesis tested in Chapter 4 was that:

Trained machine learning algorithms can provide a more effective method of detecting the ABR compared to prominent statistical detection methods, specifically with regard to sensitivity and specificity.

The work presented in Chapter 4 (Figure 4-11) showed how the presented stacked algorithm performed statistically significantly better than all of the statistical ABR detection methods evaluated, across all of the ensemble sizes evaluated. The recommendation in this thesis of comparing the performance of newly proposed machine learning algorithms with readily implementable statistical detection algorithms such as the Hotelling's T^2 test, makes comparison of the performance of algorithms used in different studies more straightforward.

A hurdle to the clinical implementation of machine learning algorithms to detect the ABR is the need for a controlled level of specificity to be achieved across a wide range of ensemble sizes. This is necessary to ensure that the false positive rate (i.e. falsely detecting a response where none is present) is stable and consistent. The work presented evaluated two methods for achieving this Figure 4-10:

1. Using a separate set of data by which to determine the detection criterion of the algorithm.
2. Using the bootstrap technique (Lv, Simpson and Bell, 2007) (Section 3.3).

Whilst the false positive rate was not controlled effectively using a separate set of data, the bootstrap technique was found to be effective in doing this. By estimating the null distribution of the output of the machine learning algorithm for each individual ensemble, the significance level of the machine learning algorithm prediction can be obtained. Overcoming this hurdle brings the field of ABR detection using machine learning one step closer towards the goal of clinical use.

From an ethical perspective, machine learning algorithms must adhere to the ethical and legal structures in place (Vayena, Blasimme and Cohen, 2018). Machine learning algorithms used in clinical practice should be 'representative of the target population' (Vayena, Blasimme and Cohen, 2018). Vayena, Blasimme and Cohen (2018) emphasise that evidence of safety and efficacy is important for machine learning algorithms used in healthcare. This should be evidenced by demonstratable good performance across all individuals for whom the algorithm will be used:

- Individuals from a variety of age groups (e.g. pre-term neonates/newborns/infants).
- Individuals with a variety of hearing levels (individuals with normal hearing, individuals with mild/moderate/severe/profound hearing losses).
- Individuals with other medical factors which may be expected to affect their EEG recordings.

7.2 Automated ABR Detection and Weighted Averaging

Weighted averaging (Elberling and Wahlgreen, 1985) has long been known to be effective at improving the SNR in the coherently weighted average. However, the literature was lacking in experiments confirming the optimal parameters for the technique. The experiment presented in Chapter 5 uses a large body of subject recorded data in combination with simulation to extensively evaluate and optimise the key parameter for weighted averaging: the block size. This work addresses a gap in the knowledge by providing an in-depth analysis on the effect of weighted averaging on the Fmp statistical detection method. Whilst research by Elberling and Wahlgreen (1985) provided anecdotal evidence, in the form of a selection of examples, that the Fsp was larger when using weighted averaging, further evidence was required to quantify the benefits that weighted averaging offers. An interesting and unexpected finding of the presented research was that the mean null Fmp value (i.e. calculated from the 'response absent' data) of the dataset used was significantly below the expected value of ~1. Additionally, when weighted averaging was applied, the mean Fmp value of 'response absent' data was found on average to increase, relative to the baseline of unweighted averaging. The cause of the mean Fmp value being below one was investigated. Whilst initially suspected to be due to an independence violation, after discussion with fellow researchers (J. Undurraga, personal communication, 2022), the low mean Fmp value was shown to be due to the finite length of the Fmp analysis window (Elberling and Don, 1984). Whilst the potential biasing effect of the analysis window length was described by Elberling and Don in 1984, this study quantifies the magnitude of the effect on the Fmp statistic and also demonstrates the linked effect that this has on the performance of ABR detection coupled with weighted averaging.

The inflation in the Fmp value caused by weighted averaging in the dataset used, had a predictable impact on the specificity of the detection test (the Fmp), increasing the false positive rate. The bootstrap technique (Lv, Simpson and Bell, 2007) was able to control the false positive rate for all of the block sizes investigated, mitigating the unexpected effects of weighted averaging on the Fmp test statistic.

A variety of methods have been proposed in the literature to estimate the background noise level within the blocks of recording epochs used for weighted averaging. A hypothesis investigated in this study was that:

ABR detection may be improved by more accurately estimating the variance of the background noise, using the 'VAR Whole Block' method, compared to the 'VAR MP' method.

The results in Chapter 5 (Figure 5-4) showed that the 'VAR Whole Block' method was able to achieve a statistically significantly higher partial ROC AUC than the 'VAR MP' method across the block sizes evaluated between 2–10 epochs. Optimising the noise estimation method, as well as the block size parameter helps to improve the performance of ABR detection algorithms. This work helps move the field forward by providing incremental gains in detection performance.

Whilst the pitfall that the analysis window length is associated with was described in the original paper on the Fsp statistic by Elberling and Don (1984), there has been little consideration of its impact on ABR analysis, let alone on ABR analysis combined with weighted averaging. The work in Chapter 5 highlights this issue and may be used to guide future recommendations for evoked potential equipment parameter settings for both the Fmp statistic and weighted averaging. The work in this study emphasises the complex interactions that pre-processing techniques such as filtering, artefact rejection, and weighted averaging have on the data and in turn the ABR detection methods applied to these data. The BSA recommend a high-pass filter setting of 30 Hz when recording the ABR in newborns, with an Fsp/Fmp analysis window length of 8–10 ms (depending on the evoked potential recording device used) (British Society of Audiology, 2019c). The results presented in Chapter 5 (30 Hz high-pass filter setting) and Appendix F (100-Hz high-pass filter setting) provide evidence that further research is required to optimise the Fsp/Fmp analysis window length and filter settings in combination, which may in turn be used to inform these recommended parameter settings.

7.3 Automated Analysis of the Diagnostic ABR using Machine Learning

The diagnostic ABR is used in the neurological evaluation of the auditory brainstem pathway. This test can be used to diagnose pathologies affecting the cochlear nerve and auditory brainstem, and also be used for surgical monitoring (Hall, 2007). The diagnostic ABR is typically interpreted visually by clinicians who will label the key structures (waves) of the waveform and then use the latencies of these waves to make clinical decisions. Algorithms which perform this task automatically can bring objectivity to the procedure and offer support to clinicians.

The aim of this study was:

To propose, train, and evaluate automated machine learning algorithms which are able to label waves I, III and V of the diagnostic ABR. Multiple state-of-the-art algorithms should be evaluated to select the best approach. The automated algorithm should also provide a confidence measure to help clinicians interpret the latency values provided. The aim was not to present a final model, ready for clinical implementation, but rather to identify promising algorithms which may then be evaluated on larger datasets reflective of the intended clinical population.

Of the machine learning algorithms evaluated in this study, the CNN-LSTM was found to perform the best, with a MAE of 0.025 ms (Table 6-14), and 95.9% of latency predictions within 0.1 ms of the label (Table 6-15). This level of performance is higher than that achieved by state-of-the-art algorithms reported in the literature (Habraken, van Gils and Cluitmans, 1993; Chen *et al.*, 2021). Whilst these results are promising, it is extremely difficult to meaningfully compare the performance of algorithms between studies. This is as a result of differences in the datasets used, study design, labelling procedures used, and outcome metrics reported.

A relatively novel feature of this study was that, as well as providing a prediction for the ABR wave latencies, a confidence measure was also produced. Aside from the study by Boston (1989), which uses a rule-based algorithm rather than a neural network, no other studies were identified in the literature review where the ABR labelling algorithm also provided a confidence measure. It is the author's firm belief that an effective confidence measure is advantageous for this type of algorithm in order to help clinicians with their interpretation of the predictions made by the wave labelling algorithm. As an audiologist, the present author is of the opinion that the provision of confidence measures which are evidenced to correlate well with the accuracy of wave latency predictions would help the adoption of objective ABR analysis algorithms by clinicians.

7.4 Limitations

7.4.1 ABR Detection using Machine Learning

A significant limitation the work presented in Chapter 4 is that the simulated dataset used was derived from a small sample of subject recorded data. This may mean that the presented results are not generalisable to the wider population. Additionally the ABR data used were recorded from adults with normal hearing. It will be important in clinical practice for any algorithm to work effectively for all individuals and this will need to be verified prior to clinical use.

While using the bootstrap technique to control the specificity level of the machine learning algorithm was found to be effective, it is also computationally expensive, taking approximately four minutes to compute a prediction compared to a fraction of a second without the bootstrap. This means that, in the presented format, the proposed algorithm is unlikely to be adoptable for clinical use.

7.4.2 Automated ABR Detection and Weighted Averaging

A limitation of the work presented in Chapter 5 is that it is based on a set of data using one particular set of pre-processing parameters. It is possible that the recommendations made in this

work may not be extrapolated directly to other datasets, especially if different pre-processing techniques/parameters, e.g. filter settings, are used. During the course of the study, it was found that the Fmp analysis window length (coupled with the chosen filter settings) had an impact on the Fmp statistic and interacted complexly with the weighted averaging technique. Additional data using different filter settings are presented in Appendix F. These additional data highlight how different recording parameters impact upon the optimal weighted averaging parameters.

As for the work in Chapter 4, the study relies on data recorded from a small number of individuals. Additionally, only one ABR template was used in the analysis. This may limit the generalisability of the presented findings. Additional work on this topic is required (see Section 7.5—Recommendations and Future Work).

7.4.3 Automated Analysis of the Diagnostic ABR using Machine Learning

A limitation of the present study is that only data recorded from individuals with normal hearing were used. This means that ABR wave latencies will have a lower variance compared to datasets containing combined ABR data recorded from individuals with normal hearing, individuals with a hearing loss, and individuals with neurological pathologies affecting the auditory nerve and/or auditory brainstem. The high performance level observed may therefore not generalise to more heterogenous datasets.

7.5 Recommendations and Future Work

7.5.1 ABR Detection using Machine Learning

The research presented in Chapter 4 represents a proof of concept, identifying a suitable machine learning approach to detecting the ABR using machine learning. Further research, e.g. a clinical study using large subject recorded datasets would be required, prior to the proposed machine learning methods being implemented in clinical practice. The machine learning algorithm would additionally need to be trained on data using the same recording settings as the data that it was intended to be used for. A relatively large amount of training data will likely be required, e.g. 900 training instances, for the machine learning model to be able to exceed the performance of optimised statistical detection methods (see the ABR detection learning curve presented in Appendix B). The dataset may be added to through use of the frequency domain bootstrap and simulation; however, it is recommended that this does not form the main portion of the dataset as it is unlikely to capture of the heterogeneity (i.e. between-subject variability) present in the

whole population, but rather extrapolate the variability from the subject recorded data already contributing to the dataset.

7.5.2 Automated ABR Detection and Weighted Averaging

As the results presented in Chapter 5 are based on one particular set of recording parameters, it is recommended that the optimisation process performed in this study be repeated using the intended clinical evoked potential equipment with the desired recording parameters, prior to implementation in clinical practice.

Whilst the work presented in Chapter 5 focusses on the Fmp due to its prominent use in clinical practice, other statistical detection methods such as Hotelling's T^2 test have been found to perform better (Chesnaye *et al.*, 2018). Further work in this area might evaluating the effects of weighted averaging on ABR detection when combined with other prominent statistical ABR detection methods such as Hotelling's T^2 test and the q-sample uniform scores test.

This study highlighted the impact of the finite-length Fmp analysis window on the Fmp statistic, which interacted with the weighted averaging technique. This research highlights the need for further working on the combined optimisation of the analysis window length, recording parameters, and weighted averaging parameters, using a large database of subject recorded data in order to guide clinical practice.

7.5.3 Automated Analysis of the Diagnostic ABR using Machine Learning

As the research presented in Chapter 6 uses data recorded only from individuals with normal hearing, further research to validate the performance of the algorithm on individuals with a wide variety of pathologies affecting their auditory brainstem pathway, and individuals with normal ABR waveforms, is necessary before this algorithm may be used in clinical practice.

7.6 End Note

Whilst the works in Chapters 4,5, and 6 have been presented separately, they all shared the common aim of improving objective analysis of the ABR. Sometimes it is easier to break challenges down into multiple components in order to address them in turn. The ABR is a widely used clinical test and is used both at near-threshold levels to assess hearing thresholds objectively and at suprathreshold levels for objective neurological analysis. Chapter 4 focussed on improving objective ABR detection using machine learning algorithms and demonstrated how a stacked ensemble could achieve higher ABR detection performance than prominent statistical detection

methods. The work in Chapter 5 focussed on optimising data pre-processing in order to improve the performance of the Fmp objective ABR detection method. Chapter 6 focussed on improving wave labelling for the suprathreshold diagnostic ABR. Here the challenge lies not with detection of the ABR, but with the accurate interpretation and labelling of the structures of the ABR waveform. The work in Chapter 6 showed how machine leaning algorithms can be trained to perform this task and demonstrated an improvement upon the performance levels reported in the literature. Throughout this thesis, developments have been made in improving the objective analysis of the ABR. It is hoped that, through further work, these approaches may be translated into tools for clinical use, which will be able to assist clinicians and benefit patients.

Appendix A Stacked Ensemble Algorithm

Figure A 1 shows the architecture of the stacked ensemble used in Chapter 4.

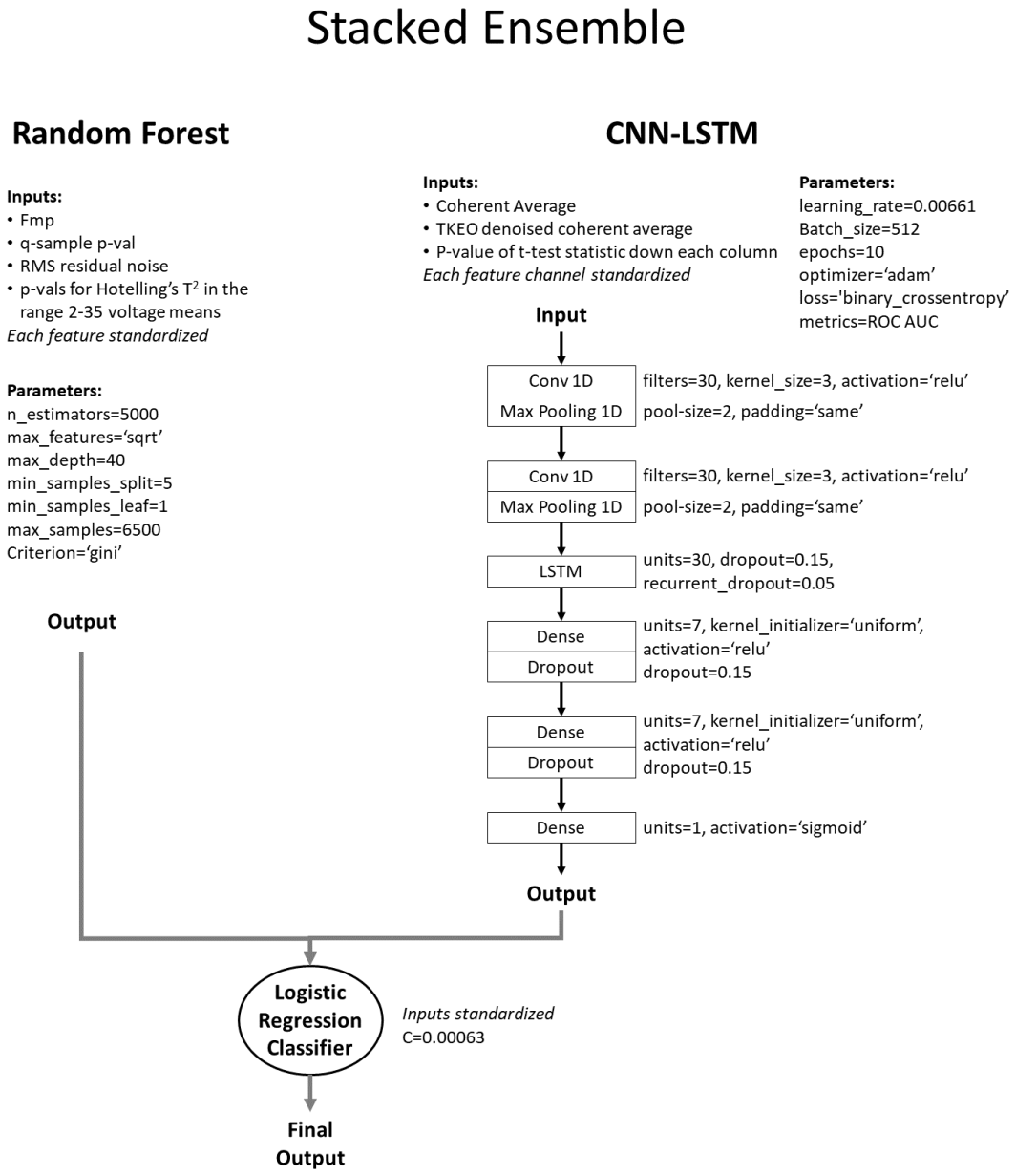

Figure A 1 Stacked ensemble algorithm architecture and optimised hyperparameters. The outputs of two base estimators (a CNN-LSTM and a random forest) are combined by a meta-estimator (a logistic regression classifier) to produce a final output prediction. The hyperparameter names in this figure are consistent with those used by the Python software libraries used to construct the algorithm: Keras (Chollet and and others, 2015) and scikit-learn (Pedregosa *et al.*, 2011). The hyperparameter values are the optimised values as obtained using the training set data (Section 4.2.4).

Figure reproduced with permission from Wolters Kluwer Health, Inc.: McKearney RM, Bell SL, Chesnaye MA, and Simpson DM. (2022) 'Auditory Brainstem Response Detection Using Machine Learning: A Comparison With Statistical Detection Methods', *Ear & Hearing*, 43(3), pp. 949–960, doi: 10.1097/AUD.0000000000001151.

Appendix B ABR Detection Learning Lurve

Obtaining large quantities of clinical data in order to train machine learning algorithms can be challenging. A learning curve analysis was performed to help understand how much data may be necessary to obtain ABR detection performance above that of the best statistical ABR detection methods. The stacked ensemble algorithm was trained on a sample of the training set data before being evaluated on the test set data (Figure A 2). The test set performance was measured using the mean ROC AUC score, with ROC AUC scores averaged across the ten different ensemble sizes evaluated (100–1,000 epochs).

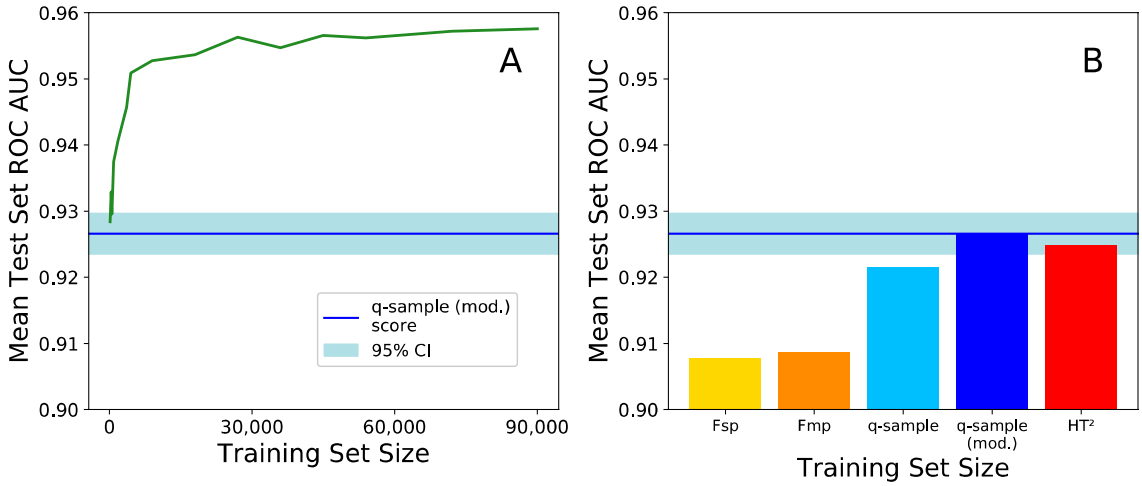

Figure A 2 Learning curve. **Plot A** shows the mean test set ABR detection performance, measured across a range of training set sizes, up to the full training set size (90,000 training instances). ABR detection performance is measured as the mean ROC AUC across all of the ensemble sizes evaluated. The performance of the best-performing statistical detection method (the modified q-sample uniform scores test) is shown as a horizontal blue line for reference. The 95% CI of this score is also shown. **Plot B** shows the mean ABR detection performance (ROC AUC) across ensemble sizes, for each of the statistical detection methods evaluated.

The ABR detection performance off the stacked ensemble remained above the upper bound of the 95% CI for the modified q-sample unform scores test for training set sizes of 900 and above. Whilst ABR detection performance began to asymptote above ROC AUC scores of 0.95, corresponding to training set sizes of 4,500 training instances and above, detection performance continued to improve with increasing training set size up to the full training set size of 90,000 training instances.

Appendix C Weighted Averaging using the Variance of the Concatenated Points in the Block

One method explored of estimating the noise level in a block of epochs is to calculate the variance of all of the concatenated points in the block. This method has the limitation of the noise variance estimate containing a bias introduced by the presence of the evoked potential signal (ABR template). It is anticipated that the low SNR of the ABR will mitigate the impact of this limitation (Sörnmo and Laguna, 2005). A simulation was performed to investigate the potential impact of the SNR of the ABR on the accuracy of the noise level estimate. The 'VAR Whole Block' method was compared with the 'VAR MP' method.

C.1 Simulation

A sine wave was added to ensembles of randomly generated noise drawn from a normal distribution. The standard deviation of this normal distribution was chosen so that 95% of the distribution would fall within ±15 μV of a mean value of zero. The ensemble sizes generated were varied, as was the SNR of the simulated data. The experiments were repeated 100 times to allow confidence intervals to be calculated.

C.2 Results

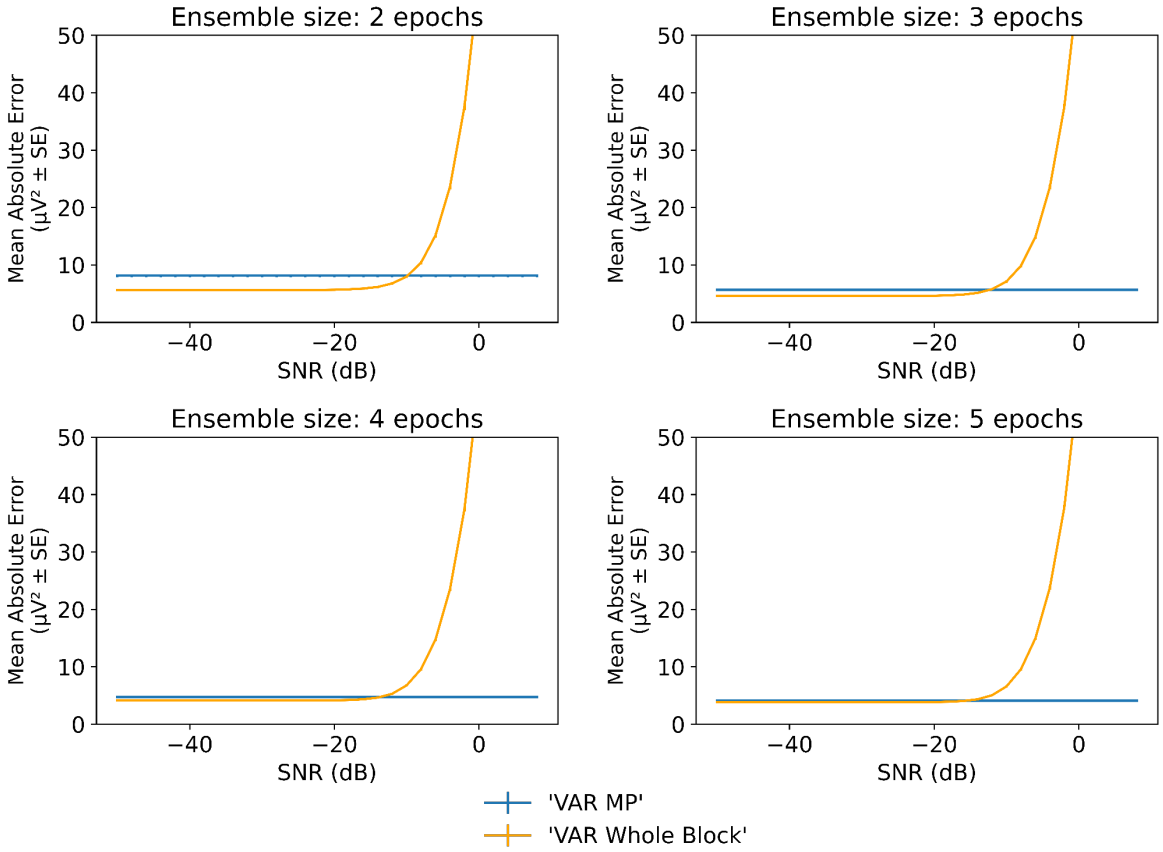

Figure A 3 A comparison of the accuracy of two noise estimation methods across ensemble sizes. The y-axis quantifies the mean absolute error between the estimate of the noise variance produced by the noise estimation method ('VAR MP'/'VAR Whole Block') and the true noise variance. The experiment was repeated 500 times to provide standard error bars. Figure reproduced without change, in accordance with the CC BY 4.0 license, from McKearney, R. M. et al. (2023) 'Optimising Weighted Averaging for Auditory Brainstem Response Detection', *Biomedical Signal Processing and Control*, 83, p. 104676. Available at: https://doi.org/10.1016/j.bspc.2023.104676.

As shown in Figure A 3, it can be seen that for all SNRs up to approximately -12 dB, the mean absolute error in the estimated variance of the noise was lower using the 'VAR Whole Block' compared to the 'VAR MP' method. For the lowest ensemble size evaluated (2 epochs), the 'VAR Whole Block' method sizeably outperformed the 'VAR MP' method in terms of the accuracy of the noise variance estimate, potentially allowing lower numbers of epochs-per-block in weighted averaging to be used. Provided the SNR of the EEG is below ~-12 dB, the mean absolute error when using the 'VAR Whole Block' method was not too large. When the SNR exceeds -12 dB, as each recording epoch becomes relatively more dominated by the variance of the ABR signal, the estimate of the variance of the noise became unstable and inaccurate. For clinical applications,

the SNR of the ABR is unlikely to exceed -12 dB and the 'VAR Whole Block' method should be a viable method to estimate the variance of the noise within blocks of epochs for weighted averaging. For higher SNR AEPs such as the CAEP, the 'VAR Whole Block' method may not be suitable/effective.

Appendix D Additional Weighted Averaging Data using the ‘VAR MP’ Method

Figure A 4 shows the sensitivity achieved using the ‘VAR MP’ method of estimating the noise levels in the blocks used for weighted averaging. These data are available for comparison with the previously presented data using the ‘VAR Whole Block’ method (Section 5.3).

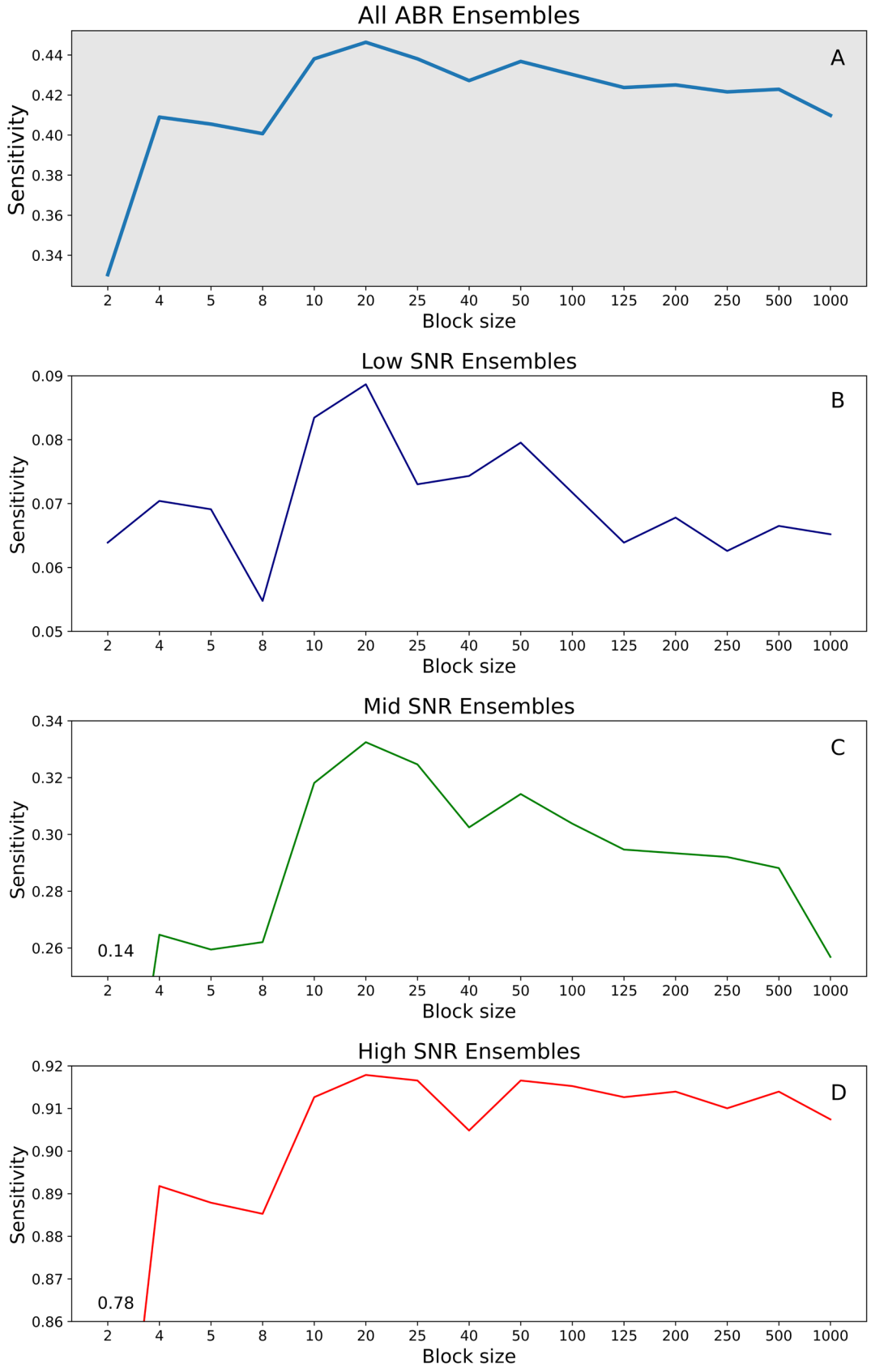

Figure A 4 The sensitivity achieved across different block sizes using the 'VAR MP' method. In order to assess the level of sensitivity fairly, the Fmp critical value was adjusted to that which achieved the desired false positive rate (0.01). Plot A shows the sensitivity level across block sizes as the proportion of all of the ABR present ensembles

correctly detected. For graphs A, B and C, the ABR present data were stratified into three evenly split groups of low- (< -32 dB), mid- (-32 to-27 dB), and high-SNR (> -27 dB) 'response present' data. The sensitivity was then calculated for that portion of the 'response present' data.

Compared to Figure 5-13 , where the 'VAR Whole Block' method was used, the data for the 'VAR MP' method showed the same main conclusion of the mid-SNR data being the sub-group to benefit most from the effects of weighted averaging. However, the peak sensitivity levels for all SNR groups were lower using the 'VAR MP' method compared to the 'VAR Whole Block' method. Peak performance for the 'VAR MP' data occurred at 20 epochs-per-block, whereas for the 'VAR Whole Block method' peak performance occurred using 10 epochs-per-block.

Figure A 5 shows the effects of weighted averaging using the 'VAR MP' method on Fmp values. Note that just as with the 'VAR Whole Block' method, the Fmp values of 'response absent' ensembles became inflated with increasingly smaller block sizes used with weighted averaging.

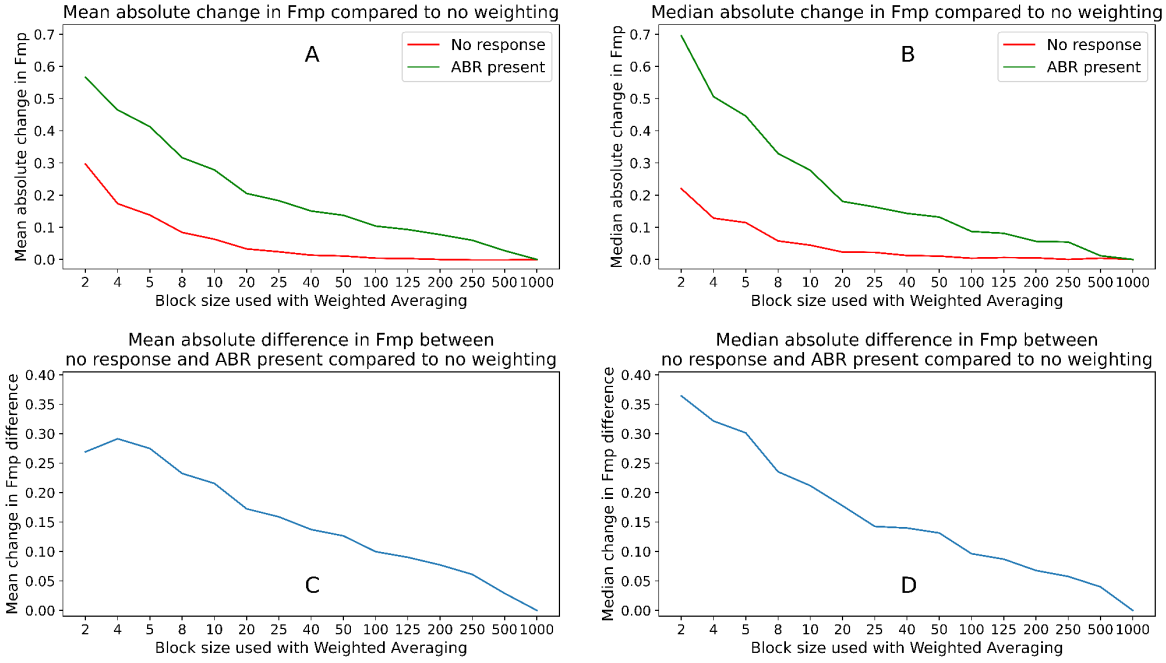

Figure A 5 Evaluation of the effects of weighted averaging using the 'VAR MP' method on Fmp values. In all four graphs, the values are presented are the absolute difference between the block size in question and a block size of 1,000, i.e. no weighting. Graphs A and C are concerned with mean values, whereas graphs B and D are concerned with median values.

Compared to Figure 5-7 in the main text, which shows the data when using the 'VAR Whole Block' method, The 'VAR MP' method generally led to a larger increase in Fmp value for the 'response present' data, but also a larger increase in Fmp value for the 'response absent' data. Despite the

difference in the Fmp values tending to be larger using the ‘VAR MP’ method compared to the ‘VAR Whole Block’ method, this did not lead to the ‘VAR MP’ method having a better detection performance or residual noise reduction (Figure 5-4, Figure 5-6, Figure A 4). This may be a result of the larger inflation of ‘response absent’ data Fmp values caused by the ‘VAR MP’ method, or less improvement in response discrimination for ensembles on the borderline of being detected or not.

Appendix E A Simulation to Explore the Effects of Serial Correlation on the Fmp Statistic

In order to investigate the reason for the low Fmp value empirically observed, a simulation was carried out to simulate decreased independence between samples by increasing the power of the low-frequency content of the simulated EEG signal. Here, 2,000 ensembles of randomly generated white Gaussian noise were generated. A low pass filter was then applied using a range of different numerator coefficients (α), with α=0 corresponding to no filter effect and α=0.99 corresponding to strong low-pass filtering. The filter equation used was:

$$y[t] = e[t] + \propto * y[t-1]$$

where $e[t]$ is the t^{th} sample of the input sequence, and $y[t]$ is the t^{th} sample of the output sequence. Each of the 2,000 ensembles were structured into 1,000 recording epochs containing 150 samples. As per the methodology used in Section 5.2, the Fmp statistic was applied to an analysis window containing samples 6–76 (inclusive). Figure A 6 shows the effect of the low-pass filter strength on the mean Fmp value.

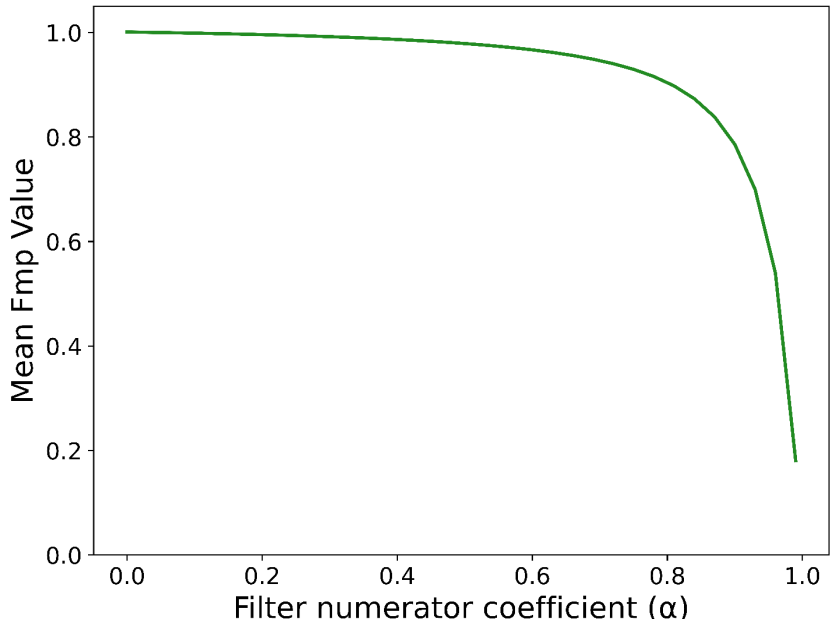

Figure A 6 The effect of serial correlation on the Fmp statistic.

Figure A 6 shows how the mean Fmp value of the ensembles decreased as increasing serial correlation was introduced by increasing the strength of the low-pass filter. It is challenging to pick apart which characteristics of the data led to the empirically observed low mean Fmp value. The *F*-test to compare two equal variances is sensitive to non-normality (Pearson, 1931; Box,

1953). It could be that non-normality is the main contributor of the empirically observed low Fmp value or that non-normality and non-independence are causing this effect in combination. Figure A 7 explores the effect of low-pass filtering on the normality of the data in both the numerator (coherent average) and denominator (single point ensemble column) of the Fsp statistic. Note that the Fsp denominator was used instead of the Fmp denominator as this allows the normality of the samples to be evaluated more readily in isolation.

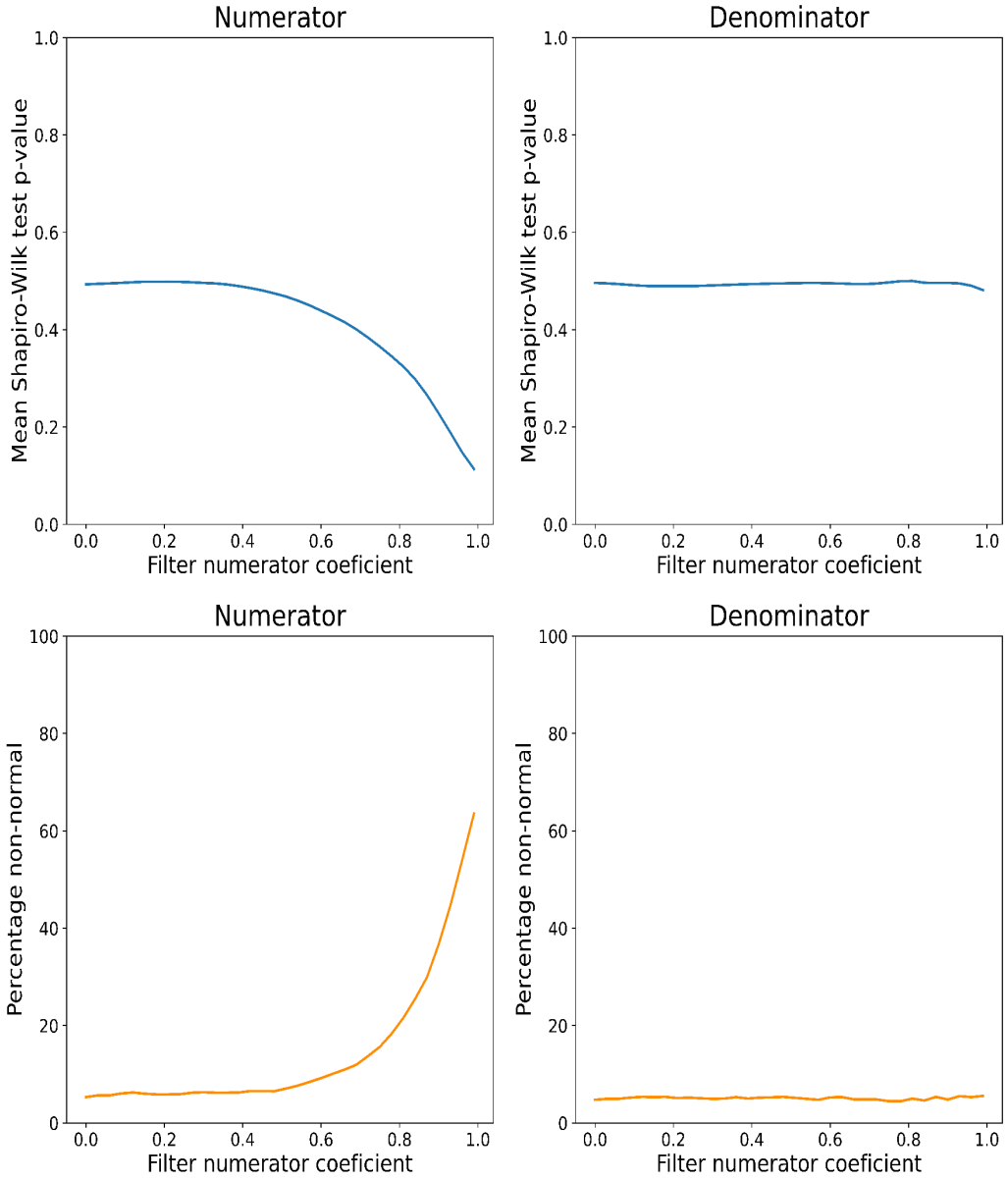

Figure A 7 The effects of sequential independence introduced through filtering on non-normality as measured using the Shapiro-Wilk test. The x-axis on all four graphs is the filter numerator coefficient, with zero corresponding to no filtering, and larger coefficient values corresponding to stronger low-pass filtering. The top two graphs (blue) show the effect of low-pass filtering on the *p* value of the Shapiro-Wilk test for normality, for samples from the numerator (coherent average) and denominator

(single point ensemble column) of the Fsp statistic. The bottom two graphs (orange) show the percentage of ensembles where the null hypothesis of normality was rejected (α set at 0.05) for both the Fsp numerator and denominator.

Figure A 7 shows how stronger low-pass filtering and therefore increased correlation between samples, led to lower *p* values using the Shapiro-Wilk test for normality and therefore a greater percentage of ensembles where the null hypothesis of samples being normally distributed was rejected.

N.B. Upon further investigation, as prompted by Dr Jaime Undurraga (J. Undurraga, personal communication, 2022), the lower-than-expected empirically obtained mean Fmp value was found to be due to the length of the analysis window (Figure 5-11) (Elberling and Don, 1984).

Appendix F Overcoming the Limitations of a finite Fmp Analysis Window Length by Raising the High-Pass Filter Setting

In order to avoid the underestimation of low-frequency spectral content as the result of a finite Fmp analysis window length, Elberling and Don (1984) suggest using an appropriate high-pass filter setting. In this appendix, results are provided whereby the data used in Chapter 5 were filtered using a raised high-pass filter setting: 100 Hz instead of the 30 Hz used previously. In addition to the EEG filter settings being changed, the already band-pass-filtered ABR template was filtered again using a high-pass filter of 100 Hz prior to being added to the no-stimulus data (this was done because the ABR database used contained already filtered ABR data).

Note that the resimulation of the data using a raised high-pass filter setting of 100 Hz, led to fewer recording epochs being rejected, and therefore a greater number of ensembles in the analysis (4,688 instead of 4,602).

Figure A 8 shows the effect of weighted averaging on the residual noise levels present in the coherently averaged waveform.

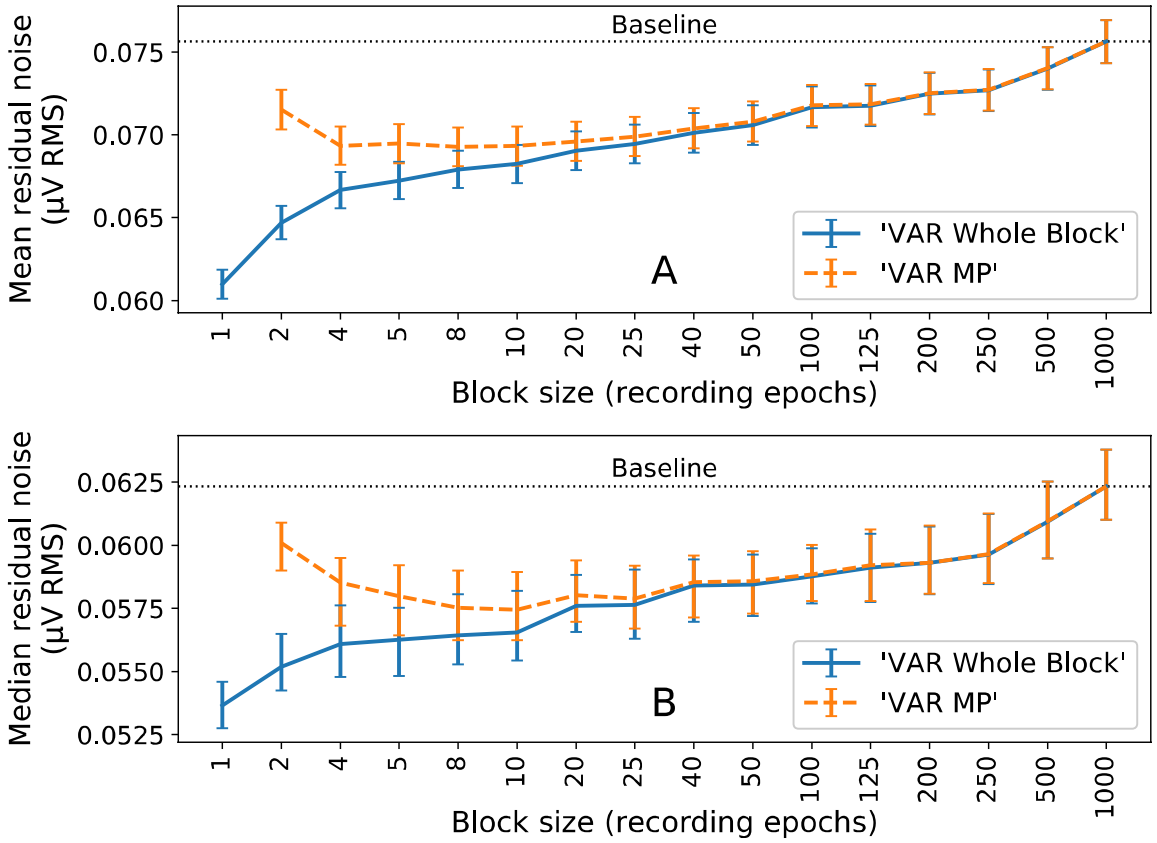

Figure A 8 Mean and median residual noise levels in the averaged waveform. The baseline represents the residual noise levels obtained using unweighted coherent averaging, i.e. 1,000 epochs-per-block. Figure reproduced without change, in accordance with the CC BY 4.0 license, from McKearney, R. M. et al. (2023) 'Optimising Weighted Averaging for Auditory Brainstem Response Detection', *Biomedical Signal Processing and Control*, 83, p. 104676. Available at: https://doi.org/10.1016/j.bspc.2023.104676.

Figure A 9 shows the ABR detection performance of the 'VAR Whole Block' method and the 'VAR MP' method.

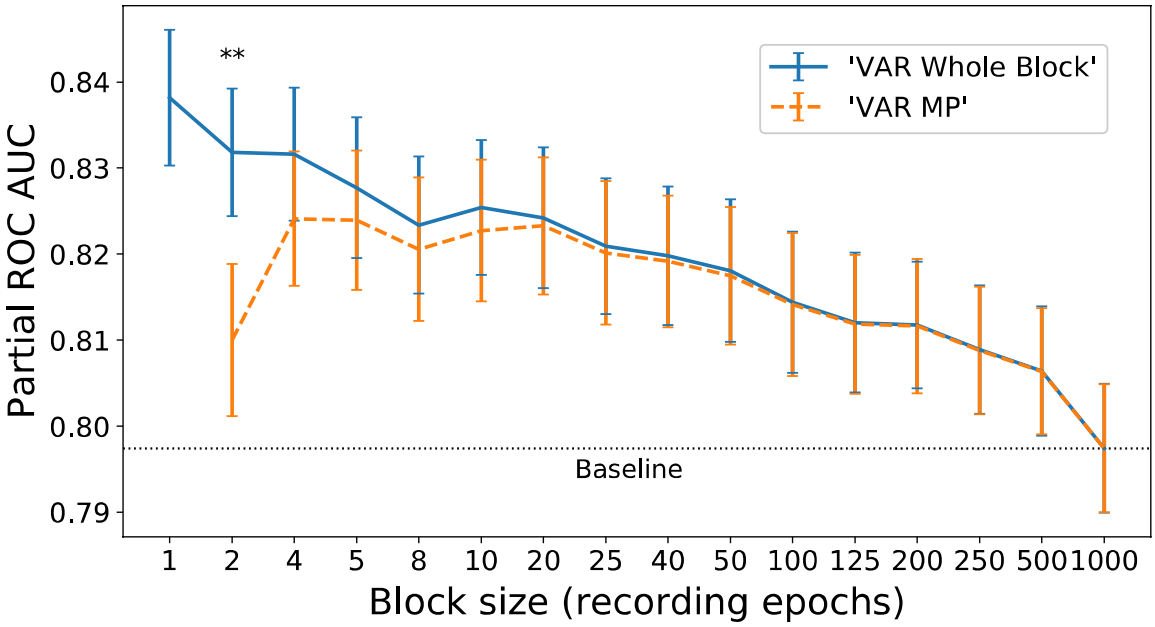

Figure A 9 Comparison of two methods for estimating the variance of the noise within each block. The evaluation metric used was the partial ROC AUC, i.e. the area under a partial region of the ROC curve, in this case the region corresponding to a false positive rate of ≤ 0.05. A higher partial ROC AUC score corresponds to a better ability to discriminate between 'response present' and 'response absent' data, over the false positive rates of interest. A double asterisk, **, indicates Bonferroni-corrected two-sided *p* value of < 0.01. Error bars represent the bootstrapped standard error of the partial ROC AUC. Figure reproduced without change, in accordance with the CC BY 4.0 license, from McKearney, R. M. et al. (2023) 'Optimising Weighted Averaging for Auditory Brainstem Response Detection', *Biomedical Signal Processing and Control*, 83, p. 104676. Available at: https://doi.org/10.1016/j.bspc.2023.104676.

Figure A 10 shows the effects of weighted averaging on the 'response present' and 'response absent' Fmp values.

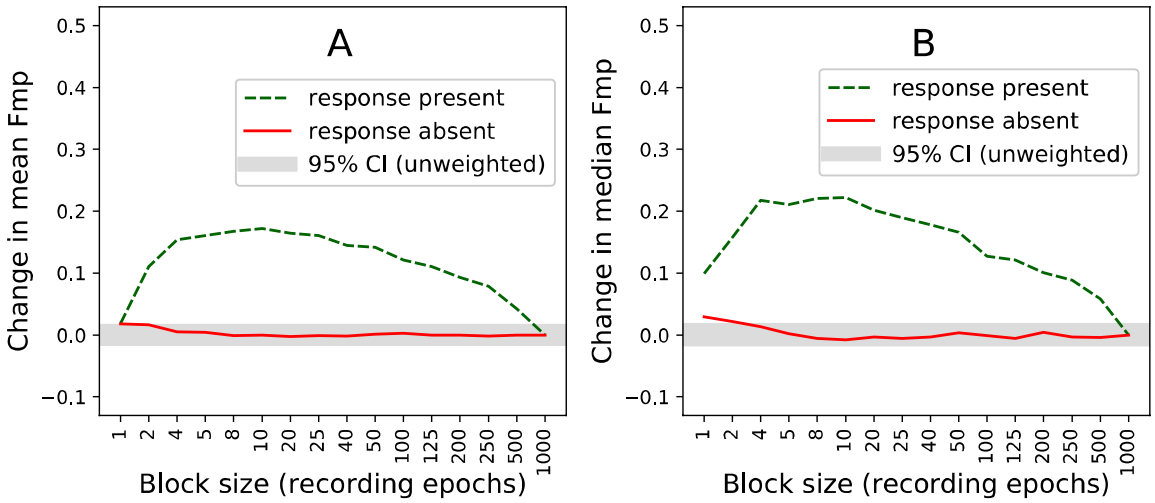

Figure A 10 Evaluation of the effects of weighted averaging on Fmp values. In both graphs, the values presented are the absolute difference between the block size in question and a block size of 1,000, i.e. no weighting. Figure reproduced without change, in accordance with the CC BY 4.0 license, from McKearney, R. M. et al. (2023) 'Optimising Weighted Averaging for Auditory Brainstem Response Detection', *Biomedical Signal Processing and Control*, 83, p. 104676. Available at: https://doi.org/10.1016/j.bspc.2023.104676.

Glossary of Terms

Coherent average..... The term used to describe the waveform produced when multiple recording epochs are averaged together. As the recording epochs are all recorded relative to the onset of the stimulus, averaging is said to be 'coherent'.

Ensemble..... An ensemble comprises a collection of N recording epochs. These all contain the same number of samples (M). An ensemble is therefore a matric of N recording epochs by M sample points.

Evoked potential ..... An electrical potential elicited by stimulation, e.g. acoustic stimuli.

Hyperparameter ..... Chosen variables which influence the machine learning process e.g. the learning rate.

k-fold cross-validation..... A resampling without replacement procedure used to select and evaluate machine learning models. Model is performance is evaluated over a number of (k) iterations, using different portions of the data each time for the evaluation (Raschka, 2020).

Recording epoch ..... A length of EEG recorded over a specified time-window, starting after the onset of a stimulus.

Training epoch ..... This is a machine learning term which refers to when the machine learning model has been trained on all of the training instances in the dataset once. A machine learning model may be trained on the full training set over multiple iterations (training epochs).

Voltage means ..... These are the mean value of multiple adjacent samples within a recording epoch, essentially combining these sample into one bin. Compression of evoked potential data into voltage means is used to improve the test performance of the Hotelling's T^2 test (Golding *et al.*, 2009).

List of References

Acir, N., Erkan, Y. and Bahtiyar, Y. A. (2013) 'Auditory brainstem response classification for threshold detection using estimated evoked potential data: Comparison with ensemble averaged data', *Neural Computing and Applications*. Springer London, 22(5), pp. 859–867. doi: 10.1007/S00521-011-0776-2.

Acir, N., Özdamar, Ö. and Güzeliş, C. (2006) 'Automatic classification of auditory brainstem responses using SVM-based feature selection algorithm for threshold detection', *Engineering Applications of Artificial Intelligence*. Pergamon, 19(2), pp. 209–218. doi: 10.1016/J.ENGAPPAI.2005.08.004.

Adam, A. *et al.* (2017) 'Improving EEG signal peak detection using feature weight learning of a neural network with random weights for eye event-related applications', *Sādhanā*, 42(5), pp. 641–653. doi: 10.1007/s12046-017-0633-9.

Ahmedt-Aristizabal, D. *et al.* (2018) 'Deep Classification of Epileptic Signals', *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*. Annu Int Conf IEEE Eng Med Biol Soc, 2018, pp. 332–335. doi: 10.1109/EMBC.2018.8512249.

Alpsan, D. (1991) 'Classification of auditory brainstem responses by human experts and backpropagation neural networks', in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society Volume 13: 1991*. IEEE, pp. 1425–1426. doi: 10.1109/IEMBS.1991.684525.

Alpsan, D. *et al.* (1994) 'Determining hearing threshold from brain stem evoked potentials. Optimizing a neural network to improve classification performance', *IEEE engineering in medicine and biology magazine*. IEEE, 13(4), pp. 465–471. doi: 10.1109/51.310986.

Anderson, N. D. (2015) 'Teaching signal detection theory with pseudoscience', *Frontiers in Psychology*. Frontiers Research Foundation, 6(JUN), p. 762. Available at: https://doi.org/10.3389/fpsyg.2015.00762 (Accessed: 15 December 2022).

Arar, S. (2019) *Use Signal Averaging to Increase the Accuracy of Your Measurements - Technical Articles*, *All About Circuits*. Available at: https://www.allaboutcircuits.com/technical-articles/use-signal-averaging-to-increase-the-accuracy-of-your-measurements/ (Accessed: 28 April 2021).

Ardila, D. *et al.* (2019) 'End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography', *Nature Medicine*. Nature Publishing Group, 25(6), pp.

954–961. doi: 10.1038/s41591-019-0447-x.

Atcherson, S. R. (2012) 'The Auditory Brainstem Response', in Atcherson, S. R. and Stoody, T. M. (eds) *Auditory Electrophysiology: A Clinical Guide*. 1st edn. New York, NY: Thieme, pp. 67–83.

Austen, S. and Lynch, C. (2009) 'Non-organic hearing loss redefined: understanding, categorizing and managing non-organic behaviour', *http://dx.doi.org/10.1080/14992020400050057*. Taylor & Francis, 43(8), pp. 449–457. doi: 10.1080/14992020400050057.

Bagatto, M. *et al.* (2005) 'Clinical Protocols for Hearing Instrument Fitting in the Desired Sensation Level Method', *Trends in Amplification*. Trends Amplif, 9(4), pp. 199–226. doi: 10.1177/108471380500900404.

Bataillou, E. *et al.* (1995) 'Weighted averaging using adaptive estimation of the weights', *Signal Processing*. Elsevier, 44(1), pp. 51–66. doi: 10.1016/0165-1684(95)00015-6.

Belue, L. M. and Bauer, K. W. (1995) 'Determining input features for multilayer perceptrons', *Neurocomputing*. Elsevier, 7(2), pp. 111–121. doi: 10.1016/0925-2312(94)E0053-T.

Benitez, J. T. *et al.* (1990) 'Evidence of central vestibulo-auditory dysfunction in atypical Cogan's syndrome: a case report - PubMed', *American Journal of Otolaryngology and Head and Neck Surgery*, 11(2), pp. 131–134.

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society*, 57(1), pp. 289–300.

Bergstra, J. and Bengio, Y. (2012) 'Random Search for Hyper-Parameter Optimization', *Journal of Machine Learning Research*, 13, pp. 281–305.

Bezerianos, A. *et al.* (1995) 'Data dependent weighted averages for recording of evoked potential signals', *Electroencephalography and Cinical Neurophysiology*. Electroencephalogr Clin Neurophysiol, 96(5), pp. 468–471. doi: 10.1016/0168-5597(95)00070-9.

Bezerra, E. D. C. *et al.* (2021) 'Dempster–Shafer Theory for Modeling and Treating Uncertainty in IoT Applications Based on Complex Event Processing', *Sensors*. Multidisciplinary Digital Publishing Institute, 21(5), p. 1863. doi: 10.3390/S21051863.

Bhargav N, M., Viswanatha, V. M. and Shailesh M L (2020) 'Wavelet Transform based Weighted Averaging Technique for Visual Evoked Potential detection', *International Journal of Advanced Science and Technology*, 29(5), pp. 3550–3561. Available at:

http://sersc.org/journals/index.php/IJAST/article/view/12047 (Accessed: 27 April 2021).

Billings, K. R. and Kenna, M. A. (1999) 'Causes of Pediatric Sensorineural Hearing Loss: Yesterday and Today', *Archives of Otolaryngology–Head & Neck Surgery*. American Medical Association, 125(5), pp. 517–521. doi: 10.1001/ARCHOTOL.125.5.517.

Bonferroni, C. E. (1936) *Teoria statistica delle classi e calcolo delle probabilità - Carlo E. Bonferroni - Google Books*. Seeber.

Boston, J. R. (1989) 'Automated Interpretation of Brainstem Auditory Evoked Potentials: A Prototype System', *IEEE Transactions on Biomedical Engineering*, 36(5), pp. 528–532.

Box, G. E. P. (1953) 'Non-Normality and Tests on Variances', *Biometrika*. Oxford Academic, 40(3–4), pp. 318–335. doi: 10.1093/BIOMET/40.3-4.318.

Bradley, A. P. and Wilson, W. J. (2004) 'On wavelet analysis of auditory evoked potentials', *Clinical Neurophysiology*. Elsevier, 115(5), pp. 1114–1128.

Bradley, A. P. and Wilson, W. J. (2005) 'Automated Analysis of the Auditory Brainstem Response Using Derivative Estimation Wavelets', *Audiology and Neurotology*, 10(1), pp. 6–21. doi: 10.1159/000081544.

Breiman, L. (1996) 'Bagging predictors', *Machine Learning*. Kluwer Academic Publishers, 24(2), pp. 123–140. doi: 10.1023/A:1018054314350.

British Society of Audiology (2019a) *Recommended Procedure: Assessment and Management of Auditory Neuropathy Spectrum Disorder (ANSD) in Young Infants*. Available at: www.thebsa.org.uk.

British Society of Audiology (2019b) *Recommended Procedure: Auditory Brainstem Response (ABR) testing for post Newborn and Adult*. Available at: www.thebsa.org.uk.

British Society of Audiology (2019c) *Recommended Procedure: Auditory Brainstem Response (ABR) testing in Babies*. British Society of Audiology. Available at: www.thebsa.org.uk.

British Society of Audiology (2021) *Guidelines for the early audiological assessment and management of babies referred from the Newborn Hearing Screening Programme*. British Society of Audiology. Available at: www.thebsa.org.uk.

Buchman, C. A. *et al.* (2006) 'Auditory neuropathy characteristics in children with cochlear nerve deficiency', *Ear and Hearing*. Ear Hear, pp. 399–408. doi: 10.1097/01.aud.0000224100.30525.ab.

Byrd, J. and Lipton, Z. C. (2019) 'What is the Effect of Importance Weighting in Deep Learning?', in *Proceedings of Machine Learning Research*, pp. 872–881.

Caceres, M. A., Sottile, F. and Spirito, M. A. (2009) 'Adaptive location tracking by Kalman filter in wireless sensor networks', *WiMob 2009 - 5th IEEE International Conference on Wireless and Mobile Computing Networking and Communication*, pp. 123–128. doi: 10.1109/WIMOB.2009.30.

Cardon, G., Campbell, J. and Sharma, A. (2012) 'Plasticity in the developing auditory cortex: Evidence from children with sensorineural hearing loss and auditory neuropathy spectrum disorder', *Journal of the American Academy of Audiology*. NIH Public Access, pp. 396–411. doi: 10.3766/jaaa.23.6.3.

Cebulla, M., Stürzebecher, E. and Elberling, C. (2006) 'Objective detection of auditory steady-state responses: Comparison of one-sample and q-sample tests', *Journal of the American Academy of Audiology*. J Am Acad Audiol, 17(2), pp. 93–103. doi: 10.3766/jaaa.17.2.3.

Cebulla, M., Stürzebecher, E. and Wernecke, K.-D. (2000) 'Comparison of several SNR estimates for objective response detection in noise', *Zeitschrift für Audiologie*, 39(1), pp. 14–22.

Chen, C. *et al.* (2021) 'Automatic Recognition of Auditory Brainstem Response Characteristic Waveform Based on Bidirectional Long Short-Term Memory', *Frontiers in Medicine*. Frontiers Media S.A., 7(613708).

Chesnaye, M. A. *et al.* (2018) 'Objective measures for detecting the auditory brainstem response: comparisons of specificity, sensitivity and detection time', *International Journal of Audiology*. Taylor and Francis Ltd, 57(6), pp. 468–478.

Chesnaye, M. A. (2019) *Optimising Objective Detection Methods for the Auditory Brainstem Response*. [Doctoral dissertation, University of Southampton].

Chesnaye, M. A. *et al.* (2021) 'Controlling test specificity for auditory evoked response detection using a frequency domain bootstrap', *Journal of Neuroscience Methods*. Elsevier, 363, p. 109352.

Chittka, L. and Brockmann, A. (2005) 'Perception Space—The Final Frontier', *PLoS Biology*. Public Library of Science, 3(4), p. e137. Available at: https://doi.org/10.1371/journal.pbio.0030137 (Accessed: 11 December 2020).

Chollet, F. (2018) *Deep Learning with Python*. Shelter Island, NY: Manning Publications Co.

Chollet, F. and and others (2015) 'Keras'. Available at: https://keras.io (Accessed: 5 February 2020).

Chui, J., Murkin, J. M. and Drosdowech, D. (2019) 'A Pilot Study of a Novel Automated Somatosensory Evoked Potential (SSEP) Monitoring Device for Detection and Prevention of Intraoperative Peripheral Nerve Injury in Total Shoulder Arthroplasty Surgery', *Journal of neurosurgical anesthesiology*, 31(3), pp. 291–298. doi: 10.1097/ANA.0000000000000505.

Cohen, M. M. *et al.* (1971) 'Auditory Evoked Response (AER): Consistency of Detection in Young Sleeping Children', *Archives of Otolaryngology*. American Medical Association, 94(3), pp. 214–219.

Cone, B. and Norrix, L. W. (2015) 'Measuring the advantage of kalman-weighted averaging for auditory brainstem response hearing evaluation in infants', *American Journal of Audiology*. American Speech-Language-Hearing Association, 24(2), pp. 153–168.

Constantine, L. L. and Lockwood, L. A. D. (1999) *Software for use : a practical guide to the models and methods of usage-centered design*. Reading, MA: Addison-Wesley Professional.

Dass, S., Holi, M. S. and Soundararajan, K. (2016) 'Classification of brainstem auditory evoked potentials using artificial neural network based on time and frequency domain features', *Journal of Clinical Engineering*. Lippincott Williams and Wilkins, 41(2), pp. 72–82. doi: 10.1097/JCE.0000000000000148.

Davenport Jr., W. B. and Root, W. L. (1987) *An Introduction to the Theory of Random Signals and Noise*. IEEE PRESS. New York, NY: IEEE PRESS.

Davey, R. *et al.* (2007) 'Auditory brainstem response classification: A hybrid model using time and frequency features', *Artificial Intelligence in Medicine*. Elsevier, 40(1), pp. 1–14. doi: 10.1016/J.ARTMED.2006.07.001.

Davila, C. E. and Mobin, M. S. (1992) 'Weighted Averaging of Evoked Potentials', *IEEE Transactions on Biomedical Engineering*. IEEE Trans Biomed Eng, 39(4), pp. 338–345. doi: 10.1109/10.126606.

Dawson, G. D. (1954) 'A summation technique for the detection of small evoked potentials', *Electroencephalography and Clinical Neurophysiology*. Elsevier, 6(C), pp. 65–84. doi: 10.1016/0013-4694(54)90007-3.

Delgado, R. E. and Özdamar, Ö. (1994) 'Automated Auditory Brainstem Response Interpretation', *IEEE Engineering in Medicine and Biology Magazine*. Institute of Electrical and Electronics Engineers Inc., 13(2), pp. 227–237. Available at: https://miami.pure.elsevier.com/en/publications/automated-auditory-brainstem-response-interpretation.

Dobie, R. A. and Wilson, M. J. (1994) 'Objective detection of 40 Hz auditory evoked potentials:

phase coherence vs. magnitude-squared coherence', *Electroencephalography and Clinical Neurophysiology/ Evoked Potentials*. Elsevier, 92(5), pp. 405–413. doi: 10.1016/0168-5597(94)90017-5.

Don, M. and Elberling, C. (1994) 'Evaluating Residual Background Noise In Human Auditory Brain-Stem Responses', *Journal of the Acoustical Society of America*, 96(5), pp. 2746–2757.

Doyle, K. J. (1999) 'Is there still a role for auditory brainstem response audiometry in the diagnosis of acoustic neuroma?', *Archives of Otolaryngology - Head and Neck Surgery*. American Medical Association, pp. 232–234. doi: 10.1001/archotol.125.2.232.

Van Drongelen, W. (2018) *Signal processing for neuroscientists*. 2nd edn, *Signal Processing for Neuroscientists*. 2nd edn. Academic Press.

Van Dun, B., Dillon, H. and Seeto, M. (2015) 'Estimating hearing thresholds in hearing-impaired adults through objective detection of cortical auditory evoked potentials', *Journal of the American Academy of Audiology*. American Academy of Audiology, 26(4), pp. 370–383. doi: 10.3766/jaaa.26.4.5.

Efron, B. and Gong, G. (1983) 'A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation', *The American Statistician*. JSTOR, 37(1), p. 48. doi: 10.2307/2685844.

Eggermont, J. J. *et al.* (1996) 'Comparison of distortion product otoacoustic emission (DPOAE) and auditory brain stem response (ABR) traveling wave delay measurements suggests frequency-specific synapse maturation', *Ear and Hearing*. Ear Hear, 17(5), pp. 386–394. doi: 10.1097/00003446-199610000-00004.

Elberling, C. (1979) 'Auditory Electrophysiology: The Use of Templates and Cross Correlation Functions in the Analysis of Brain Stem Potentials', *Scandinavian Audiology*. Taylor & Francis, 8(3), pp. 187–190.

Elberling, C. and Don, M. (1984) 'Quality estimation of averaged auditory brainstem responses.', *Scandinavian audiology*, 13(3), pp. 187–97.

Elberling, C. and Wahlgreen, O. (1985) 'Estimation of auditory brainstem response, abr, by means of bayesian inference', *Scandinavian Audiology*. Scand Audiol, 14(2), pp. 89–96.

Fan, Z. and Wang, T. (1992) 'Weighted averaging method for evoked potential: Determination of weighted coefficients', in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Institute of Electrical and Electronics Engineers Inc., pp. 2473–2474. doi: 10.1109/IEMBS.1992.5761546.

Faust, O. *et al.* (2018) 'Automated detection of atrial fibrillation using long short-term memory network with RR interval signals', *Computers in biology and medicine*. Comput Biol Med, 102, pp. 327–335. doi: 10.1016/J.COMPBIOMED.2018.07.001.

De Fauw, J. *et al.* (2018) 'Clinically applicable deep learning for diagnosis and referral in retinal disease', *Nature Medicine*. Nature Publishing Group, 24(9), pp. 1342–1350. doi: 10.1038/s41591-018-0107-6.

Fawcett, T. (2006) 'An introduction to ROC analysis', *Pattern Recognition Letters*. North-Holland, 27(8), pp. 861–874.

Feiveson, A. H. (2002) 'Power by Simulation', *The Stata Journal*. SAGE Publications: Los Angeles, CA, 2(2), pp. 107–124. doi: 10.1177/1536867X0200200201.

Felinger, A. (1998) 'Peak detection', in Felinger, A. (ed.) *Data Handling in Science and Technology - Volume 21*. Elsevier, pp. 183–190.

Fisher, R. A. (1935) *The Design of Experiments*. New York, NY: Hafner.

Fortnum, H. *et al.* (2009) 'The role of magnetic resonance imaging in the identification of suspected acoustic neuroma: a systematic review of clinical and cost effectiveness and natural history', *Health technology assessment (Winchester, England)*. Health Technol Assess, 13(18). doi: 10.3310/HTA13180.

Fotiadou, E. *et al.* (2021) 'A dilated inception CNN-LSTM network for fetal heart rate estimation', *Physiological measurement*. Physiol Meas, 42(4). doi: 10.1088/1361-6579/ABF7DB.

Freeman, D. T. (1992) 'Computer Applications in Otolaryngology: Computer Recognition of Brain Stem Auditory Evoked Potential Wave V by a Neural Network', *Annals of Otology, Rhinology & Laryngology*, 101(9), pp. 782–790. doi: 10.1177/000348949210100913.

Fridman, J. *et al.* (1982) 'Application of digital filtering and automatic peak detection to brain stem auditory evoked potential', *Electroencephalography and Clinical Neurophysiology*, 53(4), pp. 405–416. doi: 10.1016/0013-4694(82)90005-0.

Fulcher, A. *et al.* (2012) 'Listen up: Children with early identified hearing loss achieve age-appropriate speech/language outcomes by 3years-of-age', *International Journal of Pediatric Otorhinolaryngology*, 76(12), pp. 1785–1794. doi: 10.1016/j.ijporl.2012.09.001.

Galambos, R., Makeig, S. and Talmachoff, P. J. (1981) 'A 40-Hz auditory potential recorded from the human scalp', *Proceedings of the National Academy of Sciences of the United States of*

America, 78(4 II), pp. 2643–2647. doi: 10.1073/pnas.78.4.2643.

Gelfand, S. A. (2009) *Essentials of Audiology*. 3rd edn. New York, NY: Thieme.

Géron, A. (2017) *Hands-on machine learning with Scikit-Learn & TensorFlow*. 1st edn. Edited by N. Tache. Sebastopol, CA: O'Reilly.

Gerull, G., Graffunder, A. and Wernicke, M. (1996) 'Averaging evoked potentials with an improved weighting algorithm', *Scandinavian Audiology*. Taylor and Francis A.S., 25(1), pp. 21–27.

Golding, M. *et al.* (2009) 'The detection of adult cortical auditory evoked potentials (CAEPs) using an automated statistic and visual detection', *International Journal of Audiology*, 48(12), pp. 833–842. doi: 10.3109/14992020903140928.

Good, P. (2000) *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. 2nd edn. New York, NY: Springer New York (Springer Series in Statistics). doi: 10.1007/978-1-4757-3235-1.

Gorga, M. P. *et al.* (2006) 'Using a combination of click- and tone burst-evoked auditory brain stem response measurements to estimate pure-tone thresholds', *Ear and Hearing*. NIH Public Access, 27(1), pp. 60–74. doi: 10.1097/01.aud.0000194511.14740.9c.

Grönfors, T. (1993) 'Peak identification of auditory brainstem responses with multi-filters and attributed automaton', *Computer methods and programs in biomedicine*. Comput Methods Programs Biomed, 40(2), pp. 83–87.

Gu, J. *et al.* (2018) 'Recent advances in convolutional neural networks', *Pattern Recognition*, 77, pp. 354–377. doi: 10.1016/J.PATCOG.2017.10.013.

Habraken, J. B. A., van Gils, M. J. and Cluitmans, P. J. M. (1993) 'Identification of peak V in brainstem auditory evoked potentials with neural networks', *Computers in Biology and Medicine*, 23(5), pp. 369–380. doi: 10.1016/0010-4825(93)90134-M.

Haenssle, H. A. *et al.* (2018) 'Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists', *Annals of Oncology*. doi: 10.1093/annonc/mdy166.

Hall, J. W. (2007) *New handbook of auditory evoked responses*. Boston: Pearson.

Hannun, A. Y. *et al.* (2019) 'Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network', *Nature Medicine*. Nature Publishing Group, 25(1), pp. 65–69. doi: 10.1038/s41591-018-0268-3.

Hartung, J., Knapp, G. and Sinha, B. K. (2008) *Statistical Meta-Analysis with Applications*, *Statistical Meta-Analysis with Applications*. Hoboken, NJ, USA: John Wiley & Sons, Inc. (Wiley Series in Probability and Statistics). doi: 10.1002/9780470386347.

Hashimoto, I. *et al.* (1981) 'Brain-stem auditory-evoked potentials recorded directly from human brain-stem and thalamus', *Brain*. Brain, 104(4), pp. 841–859. doi: 10.1093/brain/104.4.841.

Hecox, K. and Galambos, R. (1974) 'Brain Stem Auditory Evoked Responses in Human Infants and Adults', *Archives of Otolaryngology*, 99(1), pp. 30–33. doi: 10.1001/archotol.1974.00780030034006.

Ho, T. K. (1995) 'Random decision forests', in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*. Montreal, QC, Canada: IEEE Computer Society, pp. 278–282. doi: 10.1109/ICDAR.1995.598994.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long Short-Term Memory', *Neural Computation*. MIT Press: Cambridge, MA, 9(8), pp. 1735–1780. doi: 10.1162/neco.1997.9.8.1735.

Hoke, M. *et al.* (1984) 'Weighted averaging - theory and application to electric response audiometry', *Electroencephalography and Clinical Neurophysiology*. Elsevier, 57(5), pp. 484–489.

Hotelling, H. (1931) 'The Generalization of Student's Ratio', *Annals of Mathematical Statistics*. Institute of Mathematical Statistics, 2(3), pp. 360–378. doi: 10.1214/AOMS/1177732979.

Hu, L. *et al.* (2015) 'Single-trial detection for intraoperative somatosensory evoked potentials monitoring', *Cognitive neurodynamics*, 9(6), pp. 589–601. doi: 10.1007/S11571-015-9348-Y.

Hummel, M. *et al.* (2016) 'Auditory Monitoring in Vestibular Schwannoma Surgery: Intraoperative Development and Outcome', *World Neurosurgery*. Elsevier, 96, pp. 444–453. doi: 10.1016/J.WNEU.2016.09.026.

Hunter, J. D. (2007) 'Matplotlib: A 2D Graphics Environment', *Computing in Science & Engineering*. IEEE Computer Society, 9(3), pp. 90–95. doi: 10.1109/MCSE.2007.55.

Ikawa, N., Morimoto, A. and Ashino, R. (2014) 'The detection of the relation of the stimulus intensity-latency of auditory brainstem response using optimal wavelet analysis', *International Conference on Wavelet Analysis and Pattern Recognition*. IEEE Computer Society, 2014, pp. 127–133.

Jerger, J. and Johnson, K. (1988) 'Interactions of age, gender, and sensorineural hearing loss on ABR latency', *Ear and hearing*. Ear Hear, 9(4), pp. 168–176. doi: 10.1097/00003446-198808000-

00002.

Jewett, D. L., Romano, M. N. and Williston, J. S. (1970) ‘Human auditory evoked potentials: possible brain stem components detected on the scalp.’, *Science (New York, N.Y.)*, 167(3924), pp. 1517–8.

John, M. S., Dimitrijevic, A. and Picton, T. W. (2001) ‘Weighted averaging of steady-state responses’, *Clinical Neurophysiology*. Elsevier, 112(3), pp. 555–562. doi: 10.1016/S1388-2457(01)00456-4.

Kaggle (2021) *Kaggle: Your Machine Learning and Data Science Community*. Available at: https://www.kaggle.com/ (Accessed: 10 August 2021).

Kaiser, J. F. (1990) ‘On a simple algorithm to calculate the “energy” of a signal’, in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Publ by IEEE, pp. 381–384. doi: 10.1109/icassp.1990.115702.

Kalman, R. E. (1960) ‘A new approach to linear filtering and prediction problems’, *Journal of Fluids Engineering, Transactions of the ASME*. American Society of Mechanical Engineers Digital Collection, 82(1), pp. 35–45. doi: 10.1115/1.3662552.

Kandaswamy, A. *et al.* (2004) ‘Neural classification of lung sounds using wavelet coefficients’, *Computers in Biology and Medicine*. Comput Biol Med, 34(6), pp. 523–537. doi: 10.1016/S0010-4825(03)00092-1.

Kennedy, C. R. *et al.* (2006) ‘Language Ability after Early Detection of Permanent Childhood Hearing Impairment’, *New England Journal of Medicine*. Massachusetts Medical Society, 354(20), pp. 2131–2141. doi: 10.1056/nejmoa054915.

Kenny, D. T. (1953) ‘Testing of differences between variances based on correlated variates.’, *Canadian journal of psychology*, 7(1), pp. 25–28. doi: 10.1037/H0083569.

Khodarahmi, M. and Maihami, V. (2022) ‘A Review on Kalman Filter Models’, *Archives of Computational Methods in Engineering*, 30(1), pp. 727–747. doi: 10.1007/S11831-022-09815-7/TABLES/2.

Kiang, N. Y. S. *et al.* (1986) ‘Single unit clues to cochlear mechanisms’, *Hearing Research*. Hear Res, 22(1–3), pp. 171–182. doi: 10.1016/0378-5955(86)90093-6.

Kim, J. H., Kim, C. M. and Yim, M. S. (2020) ‘An Investigation of Insider Threat Mitigation Based on EEG Signal Classification’, *Sensors 2020, Vol. 20, Page 6365*. Multidisciplinary Digital Publishing

Institute, 20(21), p. 6365. Available at: https://doi.org/10.3390/s20216365 (Accessed: 7 November 2022).

King, A. J. and Sininger, Y. S. (1992) 'Electrode Configuration for Auditory Brainstem Response Audiometry', *American Journal of Audiology*. American Speech Language Hearing Association, 1(2), pp. 63–67. doi: 10.1044/1059-0889.0102.63.

King, A. P. and Eckersley, R. J. (2019) 'Inferential Statistics V: Multiple and Multivariate Hypothesis Testing', in *Statistics for Biomedical Engineers and Scientists*. Elsevier, pp. 173–199. doi: 10.1016/b978-0-08-102939-8.00017-7.

Kirschstein, T. and Köhling, R. (2009) 'What is the source of the EEG?', *Clinical EEG and Neuroscience*. EEG and Clinical Neuroscience Society (ECNS), 40(3), pp. 146–149. doi: 10.1177/155005940904000305/ASSET/IMAGES/LARGE/10.1177_155005940904000305-FIG2.JPEG.

Klem, G. H. *et al.* (1999) 'The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology.', *Electroencephalography and clinical neurophysiology. Supplement*. Available at: https://www.scienceopen.com/document?vid=5960cfa8-7fde-441c-8592-35fdb9841499 (Accessed: 10 November 2022).

Kostorz, I. *et al.* (2013) 'Detection of waves of auditory brainstem response using IPAN99 algorithm', *Journal of Medical Informatics & Technologies*, 22(105595), pp. 219–226.

Krueger, G. P. (2006) 'Fatigue, Drowsy Decision-Making and Medical Error: Issues of Quality Health Care on JSTOR', *Journal of the Washington Academy of Sciences*, 92(2), pp. 41–60. Available at: https://www.jstor.org/stable/24531211?seq=1#metadata_info_tab_contents (Accessed: 17 November 2021).

Krumbholz, K., Hardy, A. J. and de Boer, J. (2020) 'Automated extraction of auditory brainstem response latencies and amplitudes by means of non-linear curve registration', *Computer Methods and Programs in Biomedicine*. Elsevier, 196, p. 105595. doi: 10.1016/J.CMPB.2020.105595.

Kumaragamage, C. L., Lithgow, B. J. and Moussavi, Z. K. (2016) 'Investigation of a new weighted averaging method to improve SNR of electrocochleography recordings', *IEEE Transactions on Biomedical Engineering*. IEEE Computer Society, 63(2), pp. 340–347. doi: 10.1109/TBME.2015.2457412.

LeCun, Y. *et al.* (1998) 'Gradient-based learning applied to document recognition', *Proceedings of*

the IEEE, 86(11), pp. 2278–2324. doi: 10.1109/5.726791.

Li, X., Sokolov, Y. and Kunov, H. (2002) 'System and method for processing low signal-to-noise ratio signals (U.S. Patent No. US7286983B2)'. United States, Patent.

Li, Y. *et al.* (2021) 'Deep Bayesian Gaussian processes for uncertainty estimation in electronic health records', *Scientific Reports*. Nature Publishing Group, 11(1), pp. 1–13. doi: 10.1038/s41598-021-00144-6.

Lightfoot, G. and Stevens, J. (2014) 'Effects of artefact rejection and bayesian weighted averaging on the efficiency of recording the newborn ABR', *Ear and Hearing*. Ear Hear, 35(2), pp. 213–220.

Linnebjerg, L. B., Hansen, A. E. and Møller, T. R. (2017) 'Hearing screening in newborns in the central Denmark region', *Danish Medical Journal*. Danish Medical Association, 64(4). Available at: https://pubmed.ncbi.nlm.nih.gov/28385167/ (Accessed: 25 January 2021).

Liu, W. (2001) 'The Dempster-Shafer Theory of Evidence', in *Propositional, Probabilistic and Evidential Reasoning*. Physica, Heidelberg, pp. 119–158. doi: 10.1007/978-3-7908-1811-6_6.

Luque, A. *et al.* (2019) 'The impact of class imbalance in classification performance metrics based on the binary confusion matrix', *Pattern Recognition*. Pergamon, 91, pp. 216–231. doi: 10.1016/J.PATCOG.2019.02.023.

Lütkenhöner, B., Hoke, M. and Pantev, C. (1985) 'Possibilities and limitations of weighted averaging', *Biological Cybernetics*. Springer-Verlag, 52(6), pp. 409–416.

Lv, J., Simpson, D. M. and Bell, S. L. (2007) 'Objective detection of evoked potentials using a bootstrap technique', *Medical Engineering and Physics*, 29(2), pp. 191–198.

Lyons, R. (2010) *Understanding Digital Signal Processing*. 3rd edn. Upper Saddle River, NJ: Pearson.

Maag, U. R. (1966) 'A k-sample analogue of Watson's U^2 statistic', *Biometrika*, 53(3–4), pp. 579–583.

Madden, C. *et al.* (2002) 'Clinical and Audiological Features in Auditory Neuropathy', *Archives of Otolaryngology–Head & Neck Surgery*. American Medical Association, 128(9), pp. 1026–1030. doi: 10.1001/ARCHOTOL.128.9.1026.

Madsen, S. M. K. (2010) *Accuracy of averaged auditory evoked potential amplitude and latency estimates*. [Master's thesis, Technical University of Denmark].

Madsen, S. M. K. *et al.* (2018) 'Accuracy of averaged auditory brainstem response amplitude and latency estimates', *International Journal of Audiology*. Taylor and Francis Ltd, 57(5), pp. 345–353. Available at: https://doi.org/10.1080/14992027.2017.1381770.

Mardia, K. V. (1972) *Statistics of Directional Data - 1st Edition*. 1st edn. Edited by Z. W. Birnbaum and E. Lukacs. London: Academic Press.

Martin, W. H. *et al.* (1994) 'New techniques of hearing assessment', *Otolaryngologic Clinics of North America*. Elsevier, 27(3), pp. 487–510.

McCullagh, P. *et al.* (2007) 'A comparison of supervised classification methods for auditory brainstem response determination - PubMed', *Studies in Health Technology and Informatics*, 129(2), pp. 1289–1293. Available at: https://pubmed.ncbi.nlm.nih.gov/17911922/ (Accessed: 9 August 2021).

McCulloch, W. and Pitts, W. (1943) 'A logical calculus of the ideas immanent in nervous activity', *The Bulletin of Mathematical Biophysics*. Kluwer Academic Publishers, 5(4), pp. 115–133. doi: 10.1007/BF02478259.

McKearney, R. M. *et al.* (2022) 'Auditory Brainstem Response Detection Using Machine Learning: A Comparison With Statistical Detection Methods', *Ear & Hearing*, 43(3), pp. 949–960.

McKearney, R. M. *et al.* (2023) 'Optimising Weighted Averaging for Auditory Brainstem Response Detection', *Biomedical Signal Processing and Control*, 83, p. 104676. Available at: https://doi.org/10.1016/j.bspc.2023.104676 (Accessed: 26 March 2023).

McKearney, R. M. and MacKinnon, R. C. (2019) 'Objective auditory brainstem response classification using machine learning', *International Journal of Audiology*, 58(4), pp. 224–230. doi: 10.1080/14992027.2018.1551633.

McLean, L., Scott, R. N. and Parker, P. A. (1996) 'Stimulus artifact reduction in evoked potential measurements', *Archives of Physical Medicine and Rehabilitation*. W.B. Saunders, 77(12), pp. 1286–1292. doi: 10.1016/S0003-9993(96)90194-X.

Medvedev, A. V., Agoureeva, G. I. and Murro, A. M. (2019) 'A Long Short-Term Memory neural network for the detection of epileptiform spikes and high frequency oscillations', *Scientific Reports*. Nature Research, 9(1), pp. 1–10. doi: 10.1038/s41598-019-55861-w.

Møller, A. (2006) *Hearing: Anatomy, Physiology, and Disorders of the Auditory System*. 2nd edn. Academic Press.

Møller, A. R. *et al.* (1981) 'Intracranially recorded responses from the human auditory nerve: New insights into the origin of brain stem evoked potentials (BSEPs)', *Electroencephalography and Clinical Neurophysiology*, 52(1), pp. 18–27. doi: 10.1016/0013-4694(81)90184-X.

Montaguti, M. *et al.* (2007) 'Comparative evaluation of ABR abnormalities in patients with and without neurinoma of VIII cranial nerve', *Acta Otorhinolaryngologica Italica*. Pacini Editore, 27(2), p. 68. Available at: www.ncbi.nlm.nih.gov/pmc/articles/PMC2640003/ (Accessed: 17 November 2021).

Mood, A. M., Graybill, F. A. and Boes, D. C. (1974) *Introduction to the Theory of Statistics*. 3rd edn. McGraw-Hill.

Motsch, J. F. (1987) *La dynamique temporelle du tronc cérébral: recueil, extraction et analyse optimale des potentiels évoqués auditifs du tronc cérébral*. [Doctoral dissertation, Université Paris XII].

Mühler, R. and Von Specht, H. (1999) 'Sorted averaging - Principle and application to auditory brainstem responses', *Scandinavian Audiology*, 28(3), pp. 145–149. doi: 10.1080/010503999424716.

Naftalin, L. (1981) 'Energy transduction in the cochlea', *Hearing Research*. Elsevier, 5(2–3), pp. 307–315. doi: 10.1016/0378-5955(81)90054-X.

New Zealand Ministry of Health (2016) 'Universal Newborn Hearing Screening and Early Intervention Programme'. Wellington: New Zealand Ministry of Health. Available at: https://www.nsu.govt.nz/system/files/resources/unhseip-policy-quality-standards-diagnostic-amplification-protocols-jan16.pdf.

Pal, P. *et al.* (1995) 'Early onset cerebellar ataxia with retained tendon reflexes: a clinical, electrophysiological and computed tomographic study - PubMed', *Journal of the Association of Physicians of India*, 43(9), pp. 608–613. Available at: https://pubmed.ncbi.nlm.nih.gov/8773062/ (Accessed: 17 November 2021).

Paparoditis, E. (2002) 'Frequency Domain Bootstrap for Time Series', in *Empirical Process Techniques for Dependent Data*. Birkhäuser Boston, pp. 365–381. doi: 10.1007/978-1-4612-0099-4_14.

Patel, V. L. and Kushniruk, A. W. (1998) 'Interface design for health care environments: the role of cognitive science.', *Proceedings of the AMIA Symposium*. American Medical Informatics Association, p. 29. Available at: /pmc/articles/PMC2232103/?report=abstract (Accessed: 22

March 2022).

Pearson, E. S. (1931) 'The Analysis of Variance in Cases of Non-Normal Variation', *Biometrika*. Oxford Academic, 23(1–2), pp. 114–133. doi: 10.1093/BIOMET/23.1-2.114.

Pedregosa, F. *et al.* (2011) *Scikit-learn: Machine Learning in Python*, *Journal of Machine Learning Research*. Available at: http://scikit-learn.sourceforge.net. (Accessed: 20 October 2020).

Peelle, J. E. (2016) *Human Auditory Pathway CC BY 4.0*. Available at: https://osf.io/u2gxc/ (Accessed: 19 December 2022).

Picton, T. *et al.* (1988) 'Evaluation of brain-stem auditory evoked potentials using dynamic time warping', *Electroencephalography and clinical neurophysiology*, 71(3), pp. 212–225.

Picton, T. W. *et al.* (1987) 'Reliability estimates for steady-state evoked potentials', *Electroencephalography and Clinical Neurophysiology/ Evoked Potentials*, 68(2), pp. 119–131. doi: 10.1016/0168-5597(87)90039-6.

Pimperton, H. and Kennedy, C. R. (2012) 'The impact of early identification of permanent childhood hearing impairment on speech and language outcomes', *Archives of Disease in Childhood*. BMJ Publishing Group Ltd, 97(7), pp. 648–653. doi: 10.1136/ARCHDISCHILD-2011-301501.

Pitman, E. J. G. (1937) 'Significance Tests Which May be Applied to Samples from any Populations. II. The Correlation Coefficient Test', *Supplement to the Journal of the Royal Statistical Society*. JSTOR, 4(2), p. 225. doi: 10.2307/2983647.

Pool, K. and Finitzo, T. (1989) 'Evaluation of a computer-automated program for clinical assessment of the auditory brainstem response', *Ear and Hearing*, 10(5), pp. 304–310. Available at: https://journals.lww.com/ear-hearing/Abstract/1989/10000/Evaluation_of_A_Computer_Automated_Program_for.6.aspx (Accessed: 1 December 2021).

Popescu, M. *et al.* (1999) 'Adaptive denoising and multiscale detection of the V wave in brainstem auditory evoked potentials.', *Audiology & neuro-otology*, 4(1), pp. 38–50. doi: 10.1159/000013818.

Pratt, H., Urbach, D. and Bleich, N. (1989) 'Auditory brainstem evoked potentials peak identification by finite impulse response digital filters', *Audiology*, 28(5), pp. 272–283. doi: 10.3109/00206098909081634.

Public Health England (2019) *Guidelines for surveillance and audiological referral for infants and children following newborn hearing screen*. Available at: https://www.gov.uk/government/publications/surveillance-and-audiological-referral-guidelines/guidelines-for-surveillance-and-audiological-referral-for-infants-and-children-following-newborn-hearing-screen (Accessed: 10 September 2021).

Public Health England (2020) *Newborn hearing screening programme (NHSP): care pathways for well babies*. Available at: https://www.gov.uk/government/publications/newborn-hearing-screening-care-pathways/newborn-hearing-screening-programme-nhsp-care-pathways-for-well-babies (Accessed: 25 January 2021).

Rahbar, S. *et al.* (2007) 'Auditory brainstem response classification using wavelet transform and multilayer feed-forward networks', *Proceedings of the 4th IEEE-EMBS International Summer School and Symposium on Medical Devices and Biosensors, ISSS-MDBS 2007*, pp. 128–131. doi: 10.1109/ISSMDBS.2007.4338309.

Rahne, T., von Specht, H. and Mühler, R. (2008) 'Sorted averaging-application to auditory event-related responses', *Journal of Neuroscience Methods*, 172(1), pp. 74–78. doi: 10.1016/j.jneumeth.2008.04.006.

Rao, G. M., Nandyala, S. P. and Satyanarayana, C. (2014) 'Fast Visual Object Tracking Using Modified kalman and Particle Filtering Algorithms in the Presence of Occlusions', *International Journal of Image, Graphics and Signal Processing*. MECS Publisher, 6(10), pp. 43–54. Available at: https://doi.org/10.5815/IJIGSP.2014.10.06 (Accessed: 22 December 2022).

Raschka, S. (2020) 'Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning', *arXiv:1811.12808v3*. Available at: http://arxiv.org/abs/1811.12808 (Accessed: 4 August 2020).

Raschka, S. and Mirjalili, V. (2017) *Python Machine Learning*. 2nd editio. Birmingham, UK: Packt Publishing Ltd.

Riedel, H., Granzow, M. and Kollmeier, B. (2001) 'Single-sweep-based methods to improve the quality of auditory brain stem responses', *Zeitschrift für Audiologie Audiologie*, 40(2), pp. 82–85.

Risum, A. B. and Bro, R. (2019) 'Using deep learning to evaluate peaks in chromatographic data', *Talanta*. Elsevier, 204, pp. 255–260. doi: 10.1016/J.TALANTA.2019.05.053.

Sampath, A. and Sumithira, T. R. (2022) 'Sparse based recurrent neural network long short term memory (rnn-lstm) model for the classification of ecg signals', *Applied Artificial Intelligence*, 36(1),

p. e2018183. doi: 10.1080/08839514.2021.2018183.

Sarlija, M., Jurisic, F. and Popovic, S. (2017) 'A convolutional neural network based approach to QRS detection', *International Symposium on Image and Signal Processing and Analysis*. IEEE Computer Society, pp. 121–125. doi: 10.1109/ISPA.2017.8073581.

Schimmel, H. (1967) 'The (±) reference: accuracy of estimated mean components in average response studies.', *Science*. Science, 157(784), pp. 92–94. doi: 10.1126/science.157.3784.92.

Schimmel, H., Rapin, I. and Cohen, M. M. (1974) 'Improving evoked response audiometry with special reference to the use of machine scoring', *International Journal of Audiology*, 13(1), pp. 33–65. doi: 10.3109/00206097409089335.

Schmidt, R. J. *et al.* (2001) 'The sensitivity of auditory brainstem response testing for the diagnosis of acoustic neuromas', *Archives of Otolaryngology - Head and Neck Surgery*. American Medical Association, 127(1), pp. 19–22. doi: 10.1001/archotol.127.1.19.

Schuster, M. and Paliwal, K. K. (1997) 'Bidirectional recurrent neural networks', *IEEE Transactions on Signal Processing*, 45(11), pp. 2673–2681. doi: 10.1109/78.650093.

Schwartz, M. and Shaw, L. (1975) *Signal Processing: Discrete Spectral Analysis Detection and Estimation*. New York, NY: McGraw-Hill.

Selters, W. A. and Brackmann, D. E. (1977) 'Acoustic Tumor Detection With Brain Stem Electric Response Audiometry', *Archives of Otolaryngology*. American Medical Association, 103(4), pp. 181–187. doi: 10.1001/archotol.1977.00780210037001.

Shafer, G. (1976) *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.

Shoeibi, A. *et al.* (2021) 'Automatic Diagnosis of Schizophrenia in EEG Signals Using CNN-LSTM Models', *Frontiers in Neuroinformatics*. Frontiers Media S.A., 15, pp. 1–16. doi: 10.3389/FNINF.2021.777977/BIBTEX.

Sidey-Gibbons, J. A. M. and Sidey-Gibbons, C. J. (2019) 'Machine learning in medicine: a practical introduction', *BMC Medical Research Methodology*, 19(1), p. 64. doi: 10.1186/s12874-019-0681-4.

Sohmer, H. and Feinmesser, M. (1967) 'Cochlear action potentials recorded from the external ear in man', *Annals of Otology, Rhinology & Laryngology*, 76(2), pp. 427–435. doi: 10.1177/000348946707600211.

Sörnmo, L. and Laguna, P. (2005) *Bioelectrical Signal Processing in Cardiac and Neurological Applications*, *Bioelectrical Signal Processing in Cardiac and Neurological Applications*. Burlington,

MA: Academic Press.

Souliere, C. R. *et al.* (1991) 'Sudden hearing loss as the sole manifestation of neurosarcoidosis', *Otolaryngology–head and neck surgery*, 105(3), pp. 376–381. doi: 10.1177/019459989110500305.

Stach, B. A., Westerberg, B. D. and Roberson Jr, J. B. (1998) 'Auditory disorder in central nervous system miliary tuberculosis: case report', *Journal of the American Academy of Audiology*, 9(4), pp. 305–310.

Staecker, H. and Thompson, J. (2013) 'Central Auditory System, Anatomy', in *Encyclopedia of Otolaryngology, Head and Neck Surgery*. Springer Berlin Heidelberg, pp. 376–383. doi: 10.1007/978-3-642-23499-6_536.

Stapells, D. R. (2000) 'Threshold Estimation by the Tone-Evoked Auditory Brainstem Response: A Literature Meta-Analysis Evaluation du seuil de la surdite par la methode des potentiels evoques auditifs avec stimulus tonal: meta-analyse de la litterature', *La revue d'orthophonie et d'audiologie*, 24(2), pp. 74–83. doi: 10.1080/09593330903453228.

Stegeman, D. F. *et al.* (1997) 'Near- and far-fields: Source characteristics and the conducting medium in neurophysiology', *Journal of Clinical Neurophysiology*, 14(5), pp. 429–442. doi: 10.1097/00004691-199709000-00009.

Student (1908) 'The Probable Error of a Mean', *Biometrika*. JSTOR, 6(1), p. 1. doi: 10.2307/2331554.

Stürzebecher, E. *et al.* (1996) 'Verfahren zur automatischen Hörschwellenbestimmung, insbesondere bei Neugeborenen und Kleinkindern (U.S. Patent No. 6,071,246)'. United States, Patent.

Stürzebecher, E., Cebulla, M. and Wernecke, K.-D. (1999) 'Objective Response Detection in the Frequency Domain: Comparison of Several q-Sample Tests', *Audiology and Neurotology*. Karger Publishers, 4(1), pp. 2–11. doi: 10.1159/000013815.

Subasi, A. (2007) 'EEG signal classification using wavelet feature extraction and a mixture of expert model', *Expert Systems with Applications*. Pergamon, 32(4), pp. 1084–1093. doi: 10.1016/j.eswa.2006.02.005.

Sundaramoorthy, V. *et al.* (2000) 'A computerized database of "normal" auditory brainstem responses.', *British Journal of Audiology*, 34(3), pp. 197–201. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10905453 (Accessed: 22 October 2019).

Tang, P. C. and Patel, V. L. (1994) 'Major issues in user interface design for health professional workstations: summary and recommendations', *International journal of bio-medical computing*, 34(1–4), pp. 139–148. doi: 10.1016/0020-7101(94)90017-5.

Thalmeier, D. *et al.* (2021) 'Objective hearing threshold identification from auditory brainstem response measurements using supervised and self-supervised approaches', *arXiv*, arXiv:2112.08961v1. doi: 10.48550/arxiv.2112.08961.

Tian, J., Juhola, M. and Grönfors, T. (1997) 'Latency estimation of auditory brainstem response by neural networks', *Artificial intelligence in medicine*, 10(2), pp. 115–128. doi: 10.1016/S0933-3657(97)00389-8.

Unser, M. and Aldroubi, A. (1996) 'A review of wavelets in biomedical applications', *Proceedings of the IEEE*. Institute of Electrical and Electronics Engineers Inc., 84(4), pp. 626–638. doi: 10.1109/5.488704.

Valderrama, J. T. *et al.* (2014) 'Automatic quality assessment and peak identification of auditory brainstem responses with fitted parametric peaks.', *Computer methods and programs in biomedicine*, 114(3), pp. 262–75. doi: 10.1016/j.cmpb.2014.02.015.

Valdes-Sosa, M. J. *et al.* (2009) 'Comparison of Auditory-Evoked Potential Detection Methods Using Signal Detection Theory: Comparaison des méthods de détection des potentials évoqués auditifs du tronc cérébral au moyen de la théorie de détection du signal', *International Journal of Audiology*, 26(3), pp. 166–178. doi: 10.3109/00206098709078419.

Valizadegan, H., Nguyen, Q. and Hauskrecht, M. (2013) 'Learning classification models from multiple experts', *Journal of Biomedical Informatics*. Academic Press, 46(6), pp. 1125–1135. doi: 10.1016/J.JBI.2013.08.007.

Vannier, E. *et al.* (2001) 'Computer-assisted ABR interpretation using the automatic construction of the latency-intensity curve', *International Journal of Audiology*, 40(4), pp. 191–201.

Varma, S. and Simon, R. (2006) 'Bias in error estimation when using cross-validation for model selection', *BMC Bioinformatics*, 7(1), p. 91. doi: 10.1186/1471-2105-7-91.

Vayena, E., Blasimme, A. and Cohen, I. G. (2018) 'Machine learning in medicine: Addressing ethical challenges', *PLoS Medicine*. PLOS, 15(11). doi: 10.1371/JOURNAL.PMED.1002689.

Vidler, M. and Parker, D. (2004) 'Auditory brainstem response threshold estimation: subjective threshold estimation by experienced clinicians in a computer simulation of the clinical test.', *International journal of audiology*, 43(7), pp. 417–429.

Walsh, P., Kane, N. and Butler, S. (2005) 'The clinical role of evoked potentials', *Journal of Neurology, Neurosurgery & Psychiatry*, 76(suppl 2), pp. ii16–ii22. doi: 10.1136/JNNP.2005.068130.

Walter, S. D. (2005) 'The partial area under the summary ROC curve', *Statistics in Medicine*, 24(13), pp. 2025–2040. doi: 10.1002/sim.2103.

Wilson, E. B. (1927) 'Probable Inference, the Law of Succession, and Statistical Inference', *Journal of the American Statistical Association*, 22(158), pp. 209–212. doi: 10.1080/01621459.1927.10502953.

Wilson, J. P. (1987) 'Mechanics of middle and inner ear', *British Medical Bulletin*. Oxford University Press, 43(4), pp. 821–837. doi: 10.1093/oxfordjournals.bmb.a072220.

Wolpert, D. H. (1992) 'Stacked generalization', *Neural Networks*. Pergamon, 5(2), pp. 241–259. doi: 10.1016/S0893-6080(05)80023-1.

Wolpert, D. H. and Macready, W. G. (1997) 'No free lunch theorems for optimization', *IEEE Transactions on Evolutionary Computation*, 1(1), pp. 67–82. doi: 10.1109/4235.585893.

Wong, P. K. H. and Bickford, R. G. (1980) 'Brain stem auditory evoked potentials: the use of noise estimate', *Electroencephalography and Clinical Neurophysiology*, 50(1–2), pp. 25–34. doi: 10.1016/0013-4694(80)90320-X.

Wood, S. A., Sutton, G. J. and Davis, A. C. (2015) 'Performance and characteristics of the Newborn Hearing Screening Programme in England: The first seven years', *International Journal of Audiology*, 54(6), pp. 353–358. doi: 10.3109/14992027.2014.989548.

Xu, G. *et al.* (2020) 'A One-Dimensional CNN-LSTM Model for Epileptic Seizure Recognition Using EEG Signal Analysis', *Frontiers in neuroscience*. Front Neurosci, 14, pp. 1–9. doi: 10.3389/FNINS.2020.578126.

Xu, Z. *et al.* (2021) 'BECT Spike Detection Based on Novel EEG Sequence Features and LSTM Algorithms', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29, pp. 1734–1743. doi: 10.1109/TNSRE.2021.3107142.

Yamashita, R. *et al.* (2018) 'Convolutional neural networks: an overview and application in radiology', *Insights into Imaging*, 9(4), pp. 611–629. doi: 10.1007/s13244-018-0639-9.

Yim, J. *et al.* (2020) 'Predicting conversion to wet age-related macular degeneration using deep learning', *Nature Medicine*, 26(6), pp. 892–899. doi: 10.1038/s41591-020-0867-7.

Yoshinaga-Itano, C. *et al.* (1998) 'Language of early- and later-identified children with hearing

loss', *Pediatrics*, 102(5), pp. 1161–1171. doi: 10.1542/peds.102.5.1161.

Yoshinaga-Itano, C., Coulter, D. and Thomson, V. (2001) 'Developmental outcomes of children with hearing loss born in Colorado hospitals with and without universal newborn hearing screening programs', *Seminars in Neonatology*. W.B. Saunders Ltd, 6(6), pp. 521–529. doi: 10.1053/siny.2001.0075.

Zhang, R. *et al.* (2004) 'Feature Extraction and Classification of the Auditory Brainstem Response Using Wavelet Analysis', in *Knowledge Exploration in Life Science Informatics*. Milan, Italy: Springer, Berlin, Heidelberg, pp. 169–180. doi: 10.1007/978-3-540-30478-4_15.

Zhang, Rui *et al.* (2005) 'Classification of the auditory brainstem response (ABR) using wavelet analysis and Bayesian network', in *Proceedings - IEEE Symposium on Computer-Based Medical Systems*. Shanghai, pp. 485–490. doi: 10.1109/cbms.2005.41.

Zhang, R *et al.* (2005) 'Coupling wavelet transform with bayesian network to classify auditory brainstem responses', in *IEEE Engineering in Medicine and Biology 27th Annual Conference*. Shanghai, China, pp. 7568–7571. doi: 10.1109/IEMBS.2005.1616263.

Zimmerman, D. and Zumbo, B. (1992) 'Correction for Nonindependence of Sample Observations in ANOVA F Tests', *The Journal of Experimental Education*, 60(4), pp. 367–381.

Zwislocki, J. J. (1980) 'Theory of cochlear mechanics', *Hearing Research*, pp. 171–182. doi: 10.1016/0378-5955(80)90055-6.

Bibliography

Dillon, H. (2012) *Hearing aids*. New York, NY: Thieme Medical Publishers, Incorporated.

Dobrowolski, A. *et al*. (2016) 'Classification of auditory brainstem response using wavelet decomposition and SVM network', *Biocybernetics and Biomedical Engineering*, 36(2), pp. 427–436. doi: 10.1016/J.BBE.2016.01.003.

Haboosheh, R. (2007) *Diagnostic auditory brainstem response analysis: evaluation of signal-to-noise ratio criteria using signal detection theory*. University of British Columbia. doi: 10.14288/1.0100795.

Li, M. *et al*. (2013) 'Sex and gestational age effects on auditory brainstem responses in preterm and term infants', *Early human development*, 89(1), p. 43. doi: 10.1016/J.EARLHUMDEV.2012.07.012.

Lütkenhöner, B. (2008) 'Threshold and beyond: Modeling the intensity dependence of auditory responses', *Journal of the Association for Research in Otolaryngology*. Springer-Verlag, 9(1), pp. 102–121. doi: 10.1007/s10162-007-0102-y.

Lütkenhöner, B., Klein, J.-S. and Seither-Preisler, A. (2007) 'Near-Threshold Auditory Evoked Fields and Potentials are In Line with the Weber-Fechner Law', in *Hearing – From Sensory Processing to Perception*. Springer Berlin Heidelberg, pp. 215–225. doi: 10.1007/978-3-540-73009-5_24.

Lütkenhöner, B. and Klein, J. S. (2007) 'Auditory evoked field at threshold', *Hearing Research*, 228(1–2), pp. 188–200. doi: 10.1016/j.heares.2007.02.011.

Osman, R. Al and Osman, H. Al (2021) 'On the use of machine learning for classifying auditory brainstem responses: A scoping review', *IEEE Access*. Institute of Electrical and Electronics Engineers Inc., 9, pp. 110592–110600. doi: 10.1109/ACCESS.2021.3102096.

Rushaidin, M. M. *et al*. (2012) 'Wave V detection using continuous wavelet transform of auditory brainstem response signal', in *Progress In Electromagnetics Research Symposium Proceedings*. KL, Malaysia, pp. 1889–1893.

Sininger, Y. S., Abdala, C. and Cone-Wesson, B. (1997) 'Auditory threshold sensitivity of the human neonate as measured by the auditory brainstem response', *Hearing Research*. doi: 10.1016/S0378-5955(96)00178-5.

Spivak, L. G. (1993) ‘Spectral composition of infant auditory brainstem responses: Implications for filtering’, *International Journal of Audiology*, 32(3), pp. 185–194. doi: 10.3109/00206099309072934.

Strauss, D. J. *et al*. (2004) ‘Fast detection of wave V in ABRs using a smart single sweep analysis system’, *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, 26 I, pp. 458–461.

Vannier, E., Adam, O. and Motsch, J.-F. (2002) ‘Objective detection of brainstem auditory evoked potentials with a priori information from higher presentation levels.’, *Artificial intelligence in medicine*, 25(3), pp. 283–301. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12069764 (Accessed: 22 October 2019).

Wilson, K. G. and Stelmack, R. M. (1982) ‘Power functions of loudness magnitude estimations and auditory brainstem evoked responses’, *Perception & Psychophysics*, 31(6), pp. 561–565. doi: 10.3758/BF03204188.

Wimalarathna, H. *et al*. (2021) ‘Comparison of machine learning models to classify Auditory Brainstem Responses recorded from children with Auditory Processing Disorder’, *Computer methods and programs in biomedicine*, 200(105942). doi: 10.1016/J.CMPB.2021.105942.

Wimalarathna, H. *et al*. (2022) ‘Machine learning approaches used to analyze auditory evoked responses from the human auditory brainstem: A systematic review’, *Computer methods and programs in biomedicine*, 226. doi: 10.1016/J.CMPB.2022.107118.

Zhang, H. *et al*. (2022) ‘A Robust Extraction Approach of Auditory Brainstem Response Using Adaptive Kalman Filtering Method’, *IEEE Transactions on Biomedical Engineering*. Institute of Electrical and Electronics Engineers (IEEE), pp. 1–1.