Objective Auditory Brainstem Response Classification using Machine Learning

Richard M. McKearney [1], ORCID iD 0000-0001-7030-5617, Corresponding Author

Robert C. MacKinnon [2], ORCID iD 0000-0002-6486-5578

[1] Audiology Department, Guy's and St Thomas' NHS Foundation Trust, London, UK

[2] Department of Vision and Hearing Sciences, Anglia Ruskin University, Cambridge, UK

Richard McKearney

Audiology Department

Guy's Hospital

Great Maze Pond

London

SE1 9RT

Email: richard.mckearney@nhs.net

Key words:

Auditory Brainstem Evoked Response

Classification

Supervised Machine Learning

Neural Network Models

Objective Auditory Brainstem Response Classification using Machine Learning

Abstract

OBJECTIVE: To use machine learning in the form of a deep neural network to objectively classify paired auditory brainstem response waveforms into either: 'clear response', 'inconclusive' or 'response absent'.

DESIGN: A deep convolutional neural network was trained and fine-tuned using stratified 10-fold cross-validation on 190 paired ABR waveforms. The final model was evaluated on a test set of 42 paired waveforms.

STUDY SAMPLE: The full dataset comprised 232 paired ABR waveforms recorded from eight normal-hearing individuals. The dataset was obtained from the PhysioBank database. The paired waveforms were independently labelled by two audiological scientists in order to train and evaluate the network's performance.

RESULTS: The trained neural network was able to classify paired ABR waveforms with 92.9% accuracy. The sensitivity and specificity were 92.9% and 96.4% respectively.

CONCLUSIONS: This neural network may have clinical utility in assisting clinicians with waveform classification for the purpose of hearing threshold estimation. Further evaluation on a large clinically-obtained dataset would provide further validation with regards to the clinical potential of the neural network in diagnostic adult testing, newborn testing and in automated newborn hearing screening.

Acronyms and Abbreviations

ABR	Auditory Brainstem Response
AI	Artificial Intelligence
ANN	Artificial Neural Network
CPU	Central Processing Unit
CR	Clear Response
F_{SP}	Single-point F-ratio
Inc	Inconclusive
LSTM	Long Short-Term Memory
NPV	Negative Predictive Value
PPV	Positive Predictive Value
RA	Response Absent
RAM	Random Access Memory
ROC	Receiver Operating Characteristic
SPL	Sound Pressure Level

Introduction

The Auditory Brainstem Response

The Auditory Brainstem Response (ABR) represents early components of the auditory evoked response and is typically generated in the first 10ms following the presentation of an auditory stimulus (commonly a click or tone pip). The ABR in humans was first described in 1970 by Jewett et al. The ABR has up to seven distinct vertex-positive waves which correspond to synchronous neuronal activity arising from the auditory nerve and auditory brainstem structures up to the auditory projections from the medial geniculate body (Berger & Blum, 2007). The ABR has a typical morphology in the relative amplitude and latency of these waves. Features of the waveform vary when approaching auditory threshold and are extinguished altogether when the stimulus is below threshold (Figure 1). As part of a neurological assessment, the ABR offers some site-of-lesion information according to the way in which the latency and morphology of the waves are differentially affected (Berger & Blum, 2007).

ABR testing is a clinically useful tool particularly suited to obtaining objective estimates of behavioural auditory thresholds for patients who are unable to be tested reliably using behavioural audiometric tests e.g. neonates, some adults with learning difficulties and individuals with a suspected non-organic presentation (Acır et al, 2006). Interpretation of test results is performed by visual examination of the ABR waveforms by a trained clinician, usually with the assistance of established guidelines, such as those from the Newborn Hearing Screening Programme adopted by the British Society of Audiology (Sutton & Lightfoot, 2013). It is somewhat paradoxical that, for what is typically considered an objective assessment, a significant proportion of the interpretation from which the threshold estimates

are derived, is based on the subjective interpretation of the ABR waveforms by the clinician. Interpretation of ABR waveforms therefore requires significant skill and experience. It has been shown that even amongst experienced clinicians examining the same waveforms, there can be significant variation in their interpretation and the estimated ABR threshold (Vidler & Parkert, 2004). Vidler and Parkert (2004) report that the difference between the highest and lowest estimated hearing threshold across 16 independent assessors was 40dB or greater for 9/12 sets of ABR waveforms. There is therefore significant scope to improve the accuracy of ABR assessment between clinicians by the introduction of objective classification techniques.

Objective Measures of the ABR

Some objective ABR measures exist which can be used to assist subjective classification by the clinician. These response detection techniques broadly fall into two categories. The first is syntactic comparison of the response's similarity to a model signal template, which focuses on assessing correlation (Dobie, 1993; Haboosheh, 2007). The second is comparison of the response's difference to a model background electrical noise of the waveform, which focuses on assessing amplitude or power (Dobie, 1993). Correlation methods, such as the syntactic comparison to template responses, mimic the psychophysical task performed by human interpreters. The effectiveness of these methods is limited by the heterogeneity of waveforms within and between patients (Haboosheh, 2007) and correlations requiring a high threshold in order for statistical significance to be achieved (Dobie, 1993). Amplitude or power analyses can offer more robust objective measures of ABR waveforms. Examples of these statistical detection techniques include the more widely known single-point F-ratio (F_{SP}), sometimes termed the 'quality' or 'variance' ratio (Elberling & Don, 1984; Cebulla et al, 2000), and the Standard Deviation Ratio. Both measures provide useful metrics on the level of confidence of a response being present, based on the size of the signal-to-noise ratio. However, response

classification still relies heavily on the subjective interpretation of waveform morphology by a human assessor. It is feasible that implementation of more holistic objective classification by application of machine learning techniques could further aid clinicians in interpreting ABR waveforms with a greater degree of accuracy.

Machine Learning

Machine learning is a branch of artificial intelligence (AI), which is the discipline of using machines to automate complex tasks which would normally require human intelligence to complete (Nilsson, 1971). Machine learning is the concept of enabling computers to perform a task, not through direct programming, but through learning a task from the data provided (Samuel, 1959). It has been the basis for rapid development across a range of technologies and industries, including voice recognition, translation and image recognition as well as super-human chess machines (Silver et al, 2017). Unlike traditional objective methods of classifying ABR waveforms, machine learning allows computers to select and learn the features of an ABR waveform which best relate to its correct interpretation. These features are often difficult to manually define and may include temporal and/or frequency components of the signal, or indeed a combination of different features which may not be readily observed by a human assessor visually inspecting the ABR waveform.

Previous applications of neural networks to classify ABR waveforms

Machine learning has been used previously in efforts to objectively classify ABR waveforms for the purpose of hearing threshold estimation (Alpsan, 1991; Acır et al, 2006; Davey et al, 2007). Almost all previous attempts use unreplicated ABR waveforms as the input for a neural network classifier. Many clinicians advocate the use of replicated ABR waveforms when estimating hearing thresholds, especially at stimulus levels bracketing the hearing

threshold or if no response is considered to be present (Brueggeman & Atcherson, 2012; Sutton & Lightfoot, 2013). Hobson (2016) states that 'replication of responses is essential if a correct visual interpretation is to be made'. If neural network performance is to be compared to the current gold standard of visual interpretation by a human expert, then it is considered necessary that the full information is available for the classification to be made at least for the purpose of correct labelling. The ground truth labels from which neural networks learn may therefore be incorrect if full-enough information is not available for humans to accurately define the presence or absence of a response to the gold standard. All previous artificial neural networks (ANNs) with the exception of Davey et al (2007) have been labelled by experts using only a single waveform. It is acknowledged that objective classification may be undertaken on individual waveforms, however for the purpose of comparing a classifier to the current gold-standard, paired waveforms would be required. Another limitation of previous ANNs is that they are limited to being binary classifiers trained at classifying responses into either: 'response present' or 'response absent'. For screening purposes, a binary classifier may be sufficient. However, in clinical practice, responses frequently do not definitively fall into either category and may be considered to be 'inconclusive' (Sutton & Lightfoot, 2013). For neural networks trained with a view to having clinical efficacy for diagnostic purposes in estimating hearing thresholds, it is important for them to be able to correctly classify responses which are 'inconclusive' as opposed to classifying them as 'no response'. An 'inconclusive' response at a level just below the lowest 'clear response' (e.g. at 50dBnHL) should not be considered 'response absent' as it would falsely lead to the conclusion that the threshold is definitively at 50dBnHL rather than being ≤50dBnHL as the true threshold could be anywhere at or below this level. Exemplifying the difficulties in the clinical classification of ABR waveforms, one study (Alpsan 1991) removed from the dataset all waveforms which

did not clearly fall into the categories of 'response present' or 'response absent', which would likely serve to artificially enhance the test accuracy of the ANN.

A model is presented which is trained, validated and tested using paired ABR waveforms, with each pair labelled as one of three classes: 'clear response' (CR), 'inconclusive' (Inc), or 'response absent' (RA), using the response decision categories as described by Sutton and Lightfoot (2013). To the best of the authors' knowledge, this is the first time that a deep convolutional neural network used to classify ABR waveforms has been presented in the literature.

Methods

Data

Data used to train the neural network were obtained from the PhysioBank database, (Goldberger et al, 2000) which contains ABR data from eight normal-hearing participants (four female, four male; age range 19-31). The participants had audiometric thresholds ≤15dBHL at octave frequencies spanning 250Hz – 8kHz. The data were contributed by Silva & Epstein (2010), and comprised 232 paired ABR waveforms recorded across a range of stimulus levels ranging from 5dB below the participant's threshold up to 100dB peak-equivalent SPL. The stimuli used were 1kHz and 4kHz tone pips with a 2-cycle rise/fall time with no plateau. The stimulus rate was 23.97/s. The recorded signal was band-pass filtered, with the high pass filter set at 30Hz and the low pass filter set at 3kHz. The artefact rejection level was set at ±50μV. The number of recording epochs per averaged waveform was approximately two thousand. The paired ABR waveforms were classified independently by two clinical scientists (the present authors) into one of three classes: 'clear response' (CR), 'response absent' (RA) or 'inconclusive' (Inc). The decision criteria used were based on those

described by Sutton and Lightfoot (2013). In summary, a 'clear response' is present when there is 'a high degree of correlation between waveforms', the response size is ≥40nV and ≥3 times the size of the background noise level. A response is deemed to be 'absent' if the waveforms are appropriately flat with no evidence of a response and the average noise is ≤25nV. A response is considered 'inconclusive' if neither the criteria for a 'clear response' or 'response absent' are met (Sutton and Lightfoot, 2013). For those cases where classification differed between assessors, the result was deliberated and a definitive conclusion reached by consensus. Once classified, the data were pre-processed by scaling it into the range 0-1 for use by the neural network. A recording window of 1.5-12ms was used to train and test the neural network (the first 1.5ms were omitted to avoid stimulus artefact from affecting the machine learning process, sometimes called 'stimulus blocking'). The sample rate used was 48kHz. The input to the neural network was therefore a scaled raw feature vector consisting of two channels of 504 data points.

The data were split into a training set with paired ABR waveforms derived from six participants (n=190) and a test set with data from two participants (n=42). There was no overlap of participants between the training set and the test set in order to provide a truer test of the generalisability of the model. The training set was used to train the neural network to learn from the features of paired ABR waveforms in order to make predictions regarding the class which the waveforms belonged to (CR/RA/Inc). The test set was used to evaluate the final generalisability of the neural network in classifying paired ABR waveforms which it had previously not seen.

Neural Network

The neural network was constructed in Python using Keras with a backend of TensorFlow (Keras, 2018). A Core i3 CPU processor was used with 16GB RAM.

Model construction and hyperparameter fine-tuning was conducted using stratified k-fold cross-validation (Stone, 1974; Wolpert & Macready, 1997), with k=10 folds applied to the training set. For each of the ten folds, the validation set for each fold was initially set aside, followed by the training set of the fold undergoing synthetic minority oversampling (Chawla et al, 2002). Synthetic minority oversampling helps prevent the model from being biased toward predicting an over-represented class (Inc in this case) by synthetically producing 'new' data points of the under-represented classes based on the characteristics of data from those classes within the training set. For the RA data within the training portion of each fold, a 10.5ms sample from the second half of the recording (20-30.5ms) was used to upsample the RA class so that nRA = nInc (the majority class). In order to balance the remaining under-represented CR class, random training instances from the training portion of each fold were selected from which to synthetically produce 'new' training instances. This was done by adding an amount of randomly-generated Gaussian noise to the waveforms whilst ensuring that the noise added did not cause the class of the training instance to be altered (Li & Liu, 2016). For each fold, the validation set was used to evaluate the performance of the model as a classifier. The 10-fold stratified cross-validation evaluation metrics, as shown in Table 1, were used to evaluate model performance during the process of model architecture construction and hyperparameter fine-tuning. Hyperparameters were set using a combination of manual and grid search.

The final model architecture is shown in Figure 2. The model combines convolutional, pooling and fully-connected layers. Convolutional layers draw inspiration from the structure

of the visual cortex (LeCun et al, 1999). Here, lower-level receptive fields combine outputs into higher-level neurones which therefore have a larger combined receptive field able to detect more complex patterns. The wavelet transform is a feature extraction signal processing technique and has been applied to ABR waveforms for the purpose of feature selection (Acır et al, 2006). The wavelet transform is applied convolutionally using a defined mathematical function (the wavelet). Rather than using an *a priori* defined function, a convolutional layer in a neural network applies a similar technique to the wavelet transform, although the kernel convolving the input signal learns the best mathematical function to apply to the signal in order to maximise performance in its classification.

In the final model, all three convolutional layers contained 50 filters. The first and third convolutional layers utilised a kernel size of 9, with a kernel size of 7 used in the second convolutional layer. The hidden dense (fully-connected) layer contained 30 neurons. The output layer contained 3 neurons to equal the number of classes. To prevent overfitting to the training data, regularisation in the form of dropout was applied to the penultimate fully-connected layer (Srivastava et al, 2014). He initialisation was used to initialise the weights in the network (He et al, 2015).

Once the fine-tuning process using stratified 10-fold cross-validation was completed, the best performing model was selected and trained on the full training set, with synthetic minority oversampling. The hitherto unseen test set was then used to evaluate the final performance of the model.

Statistical Analysis

The calculation of evaluation metrics for the performance of the final model was conducted using a one-vs-rest strategy applied separately for each class. The F_1 score was calculated, which provides a harmonic mean of the sensitivity (recall) and positive predictive value (precision) (Chinchor & Nancy, 1992). The Receiver Operating Characteristic (ROC) analyses, including Area Under the Curve (AUC), were performed separately for each class by varying the cut-off applied to the probabilistic output of the neural network for the class in question, using MedCalc statistical software. The micro-averaged ROC AUC was calculated using Scikit-Learn.

Ethics

The study was granted Health Research Authority approval and was hosted and sponsored by Guy's and St Thomas' NHS Foundation Trust, London, UK.

Results

The 232 paired ABR waveforms were classified into: CR (n=60), RA (n=51), and Inc (n=121). These were used as the ground truth labels with which to train and test the model. There was agreement between the two assessors for 94.8% of the paired waveforms, with consensus reached after deliberation of the remainder.

Model Selection

The performance of different neural network models as evaluated using stratified k-fold cross-validation during the model construction process is shown in Table 1. The best performing model on the validation data was a deep convolutional neural network. The mean accuracy score for 10-fold cross-validation of the best performing model was 82.0% (±5.9% SD). The evaluation metrics across each of the ten folds is shown in Table 2. The training and

validation accuracy of the model over each epoch is shown in Figure 3. The model converged at approximately 1,000 training epochs and began to overfit beyond this point.

Evaluation of the Final Model on the Test Set

The best performing model on the basis of its performance on the validation data was selected to be evaluated using the test set. The total computational time to train the final model was 16 minutes and 42 seconds. The final model accuracy on the test set was 92.9%. The micro-averaged sensitivity and specificity across classes was 92.9% and 96.4% respectively. The micro-averaged F_1 score was 0.929 and Cohen's $\kappa = 0.893$ (95% CI, 0.776 to 1.000). Table 3 shows the evaluation metrics for each class. Figure 4 shows the confusion matrix of results.

Receiver Operating Characteristic Analysis

The ROC curves for each class are shown in Figure 5. The micro-averaged ROC AUC across all three classes was 0.946.

Discussion

The present study reports the successful application of a deep convolutional neural network to classify paired ABR waveforms as either CR, RA or Inc with 92.9% accuracy, a sensitivity of 92.9% and a mean specificity of 96.4%. The deep convolutional neural network performed well at classifying paired ABR waveforms from normal-hearing individuals and this strong performance suggests that a deep convolutional neural network may have potential clinical utility in aiding clinicians to interpret ABR waveforms. This could lead to more consistent waveform interpretation between clinicians and potentially greater accuracy in establishing electrophysiological hearing thresholds. The performance of this model in the clinical setting is yet to be demonstrated and a high level of performance in classifying waveforms from both

normal-hearing and hearing impaired individuals would need to be demonstrated in order for the model to be of greatest clinical utility. This type of deep learning model may also have value as a classifier within automated hearing screening devices used as part of a Newborn Hearing Screening Programme (e.g. NHSP, UK).

The 10-fold cross-validation accuracy was lower than the final performance of the model on the test set. This can occur because the final model was trained on the full training set with no validation sets set aside. As a result there is more data available for the final model to learn from. Alternatively, chance variation could have led to clearer clear responses and more conclusively inconclusive waveforms residing in the test set rather than the training set. A larger amount of data would likely boost the model's performance and generalisability as well as minimise the effects of chance variation. When constructing and fine-tuning the model using stratified k-fold cross-validation, a variety of model architectures were considered (Wolpert & Macready, 1997). Various combinations of: convolutional layers, pooling layers, recurrent layers using Long Short-Term Memory (LSTM) units, bidirectional recurrent layers, and fully-connected neural network layers were considered. The best performance on k-fold cross-validation was on the model as depicted in Figure 2. Convolutional neural networks have been used to excellent effect in classifying electrocardiogram (ECG) waveforms (Rajpurkar et al, 2017). Given the model's good performance in ABR waveform classification, similar architectures may prove favourable in the classification in other evoked potential measurements as well as that of other waveforms. The convolution operation is translation invariant and is therefore useful in detecting events which may occur at different points in time.

Related Work

Previous efforts have been made to classify ABR waveforms using machine learning algorithms. Alpsan (1991), with 285 ABR recording available, used a feed-forward multi-layered ANN to classify single ABR waveforms into either 'response' or 'no response'. The ANN was able to correctly classify 74.9% of waveforms. Responses which were deemed to not to fall neatly into one or other category were discarded and not used to train or test the ANN. These waveforms would represent the cases which clinicians have the most difficulty in interpreting and to omit these limits the usefulness of the network and over-estimates the generalisable accuracy of the model in relation to previously unseen data. Additionally, in clinical practice, it is typical for pairs of ABR waveforms to be used in order to judge the presence or absence of a response, especially for difficult to assess waveforms, in order to evaluate the noise levels and repeatability of any perceived response. Indeed, usage of replicated waveforms has been described as essential to their visual interpretation (Hobson, 2016).

Acır et al (2006) identify that it is difficult to compare the performance of automated classification models due to variation in the data, classification techniques and the evaluation metrics used. Other previous neural networks tested are binary classifiers and therefore have a minimum expected ('chance') classification accuracy of 50% on the basis of having a balanced training set. As the classifier presented in this study is a multi-class classifier (3 classes), calculation of Cohen's kappa may provide a fairer comparison metric (Cohen, 1960). This measure of inter-rater agreement takes into account the proportion of each prediction which is expected to occur due to chance as well as any class imbalance in the dataset. Cohen's kappa for Alpsan's (1991) artificial neural network was back calculated to be $\kappa = 0.271$. This level of agreement would only be considered 'fair' (Landis & Koch, 1977).

Davey et al (2007) produced a hybrid model, combining ANNs with a C5.0 decision tree, obtaining a mean classification accuracy of 85.0% over six test sets. The mean Cohen's kappa over six test sets for this binary classifier model was back calculated to be $\kappa = 0.680$, which would be considered a 'substantial' level of agreement between the correct class and the prediction of the model. This is one of the only reported models that benefits from the use of replicated waveforms in the classification process.

Acır et al (2006) trained a support vector machine using one of three feature vectors to classify single averaged ABR waveforms into 'response present' or 'no-response', achieving a peak accuracy of 97.7% on the second feature set they used (discrete cosine transform coefficients). Cohen's kappa for the second feature set is back calculated as 0.945, indicating 'almost perfect agreement' (Landis & Koch, 1977). However, a potential limitation of this study is that the iterative process used to hone the classification features appears to use data from the test set as the validation for the purpose of feature selection. In relation to feature extraction and selection the authors write "the classification performance of the system is tested by using testing data and stored for determining the best features". If this is the case, the reported performance measures may not reflect the generalisability of the model in classifying unseen ABR data, but rather an ability of the model to selectively overfit to the test set during feature extraction/training. Additionally, it is not reported whether there was an overlap in data contribution form participants into both the training and the test which may serve to reduce the generalisability of the model.

The performance of the presented model compares very favourably to the level achieved by other researchers, despite the inclusion of an additional class making it a multi-class classifier instead of a simpler binary classifier. Additionally, it is one of the only models to use paired

instead of individual waveforms. The accuracy of the binary classifier presented by Acır et al (2006) is slightly better than that of the model presented, but has one fewer class and uses unpaired waveforms. The model presented is the first deep convolutional neural network used to classify ABR waveforms and achieves a high degree of accuracy, sensitivity and specificity.

Limitations

Large portions of the dataset contained a significant amount of noise leading to a large number of 'inconclusive' instances. This was compensated for via synthetic minority oversampling to balance the training set during training. The scales for the ABR data from the database were not in the correct order of magnitude for the physiological signal in question so a corrective amplitude multiplication was applied equally to all values for the purpose of labelling. This is not necessarily a limitation however as both the training and test sets were on the same scale and therefore the validity of the model as a classifier is not affected. The chief limitations are the relatively small size of the dataset, which has been discussed previously, and the need for further work using clinical data. The use of a larger dataset including waveforms from normal-hearing and hearing impaired individuals is required to further validate the use of this machine learning model for the purpose of generalisable, objective ABR waveform classification.

It is acknowledged that even in the pursuit of objectivity, the performance of the neural network is limited by the quality of ground truth labelling by human assessors which is known to be variable (Vidler and Parkert 2004). The present study utilised the judgement of two clinical scientists to provide the ground truth labelling, drawing upon elements of the Delphi method on the basis that group judgements tend to be more accurate than those made

on an individual basis. The present study provides evidence that a deep convolutional neural network is able to learn from the decision making process of humans to classify ABR waveforms to an accuracy of 92.9%. If trained by a panel of world-leading experts, a deep convolutional neural network would be expected, on the basis of the presented findings, to be able to learn from decision-making process of these experienced clinicians and could be used to assist those clinicians with less experience in ABR waveform interpretation.

Future Research

A significant limit on what the proposed model is able to achieve is the size of the dataset available. Collaboration between multiple sites could lead to a significant combined dataset. The authors welcome correspondence from interested collaborators. If a neural network was trained with sufficient data labelled by a panel of experts, then the predictions made by this model could theoretically reflect those of the expert panel. This algorithm could in turn exist as a module within evoked-potential recording software, providing clinicians with real-time assistance in ABR classification. The output of the neural network can be presented probabilistically with a probability value outputted for each of the three classes. This could be used as a confidence measure by clinicians. Given enough data, machine learning models have been able to match or exceed the performance of humans in a variety of complex tasks (Rajpurkar et al, 2017; He et al, 2015; Haenssle et al, 2018).

Conclusions

This study is the first reported application of a deep convolutional neural network used to classify paired ABR waveforms. It achieved a high degree of accuracy (92.9%), sensitivity (92.9%) and specificity (96.4%). Additionally, this model is the first to objectively classify pairs of ABR waveforms into either: ‘clear response’, ‘inconclusive’, or ‘response absent’,

having been trained on 190 paired waveforms labelled by two audiological scientists. The use of such a model in the future has the potential to provide real-time assistance to clinicians in interpreting ABR waveforms and lead to greater accuracy in objective threshold determination, improved automated ABR screening and reduced variation in interpretation between clinicians.

Word count = 5,021 (includes abstract, tables and figure legends)

Acknowledgements

The authors are grateful to Ikaro Silva and Michael Epstein for contributing their ABR dataset to the PhysioBank database. We would also like to thank Paul Solomon for his enthusiasm and advice regarding machine learning.

Declaration of Interest

The authors report no conflicts of interest.

References

Acır, N., Özdamar, Ö., Güzeliş, C., 2006. Automatic classification of auditory brainstem responses using SVM-based feature selection algorithm for threshold detection. *Eng. Appl. Artif. Intell.*, 19(2), p.209–218.

Alpsan, D., 1991. Classification Of Auditory Brainstem Responses By Human Experts And Backipropagation 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.

Berger, J.R., Blum, A.S., 2007. Brainstem Auditory Evoked Potentials. In *The Clinical Neurophysiology Primer*. Totowa, NJ: Humana Press, pp. 475–484.

Brueggeman, P.M., Atcherson, S.R., 2012. Threshold Estimation Using the Auditory Brainstem Response. In S. R. Atcherson & T. M. Stoody, eds. *Auditory Electrophysiology: A Clinical Guide*. New York, NY: Thieme, pp. 203–219.

Cebulla, M., Stürzebecher, E., Wernecke, K.D., 2000. Objective detection of auditory brainstem potentials: comparison of statistical tests in the time and frequency domains. *Scand. Audiol.*, 29(1), p.44–51.

Chawla, N. V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.*, 16, p.321–357.

Chinchor, N., Nancy, 1992. MUC-4 evaluation metrics. In *Proceedings of the 4th conference on Message understanding - MUC4 '92*. Morristown, NJ, USA: Association for Computational Linguistics, p. 22.

Cochran, W.G., 1950. The comparison of percentages in matched samples. *Biometrika*, 37(3–

4), p.256–66.

Cohen, J., 1960. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.*, 20(1), p.37–46.

Davey, R., McCullagh, P., Lightbody, G., McAllister, G., 2007. Auditory brainstem response classification: A hybrid model using time and frequency features. *Artif. Intell. Med.*, 40(1), p.1–14.

Dobie, R.A., 1993. Objective response detection. *Ear Hear.*, 14(1), p.31–5.

Elberling, C., Don, M., 1984. Quality estimation of averaged auditory brainstem responses. *Scand. Audiol.*, 13(3), p.187–97.

Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., et al, 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23), p.E215-20.

Haboosheh, R., 2007. *Diagnostic auditory brainstem response analysis : evaluation of signal-to-noise ratio criteria using signal detection theory*. University of British Columbia. Available at: https://open.library.ubc.ca/cIRcle/collections/ubctheses/831/items/1.0100795#share [Accessed August 8, 2018].

Haenssle, H.A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., et al, 2018. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* Available at: https://academic.oup.com/annonc/advance-article/doi/10.1093/annonc/mdy166/5004443 [Accessed July 5, 2018].

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving Deep into Rectifiers: Surpassing Human-

Level Performance on ImageNet Classification. Available at: http://arxiv.org/abs/1502.01852 [Accessed June 16, 2018].

Hobson, T., 2016. *Audiology Diagnostic Assessment Protocol*, Queensland, Australia. Available at: https://www.childrens.health.qld.gov.au/wp-content/uploads/PDF/healthy-hearing/hh-audiology-protocol.pdf.

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

Keras, 2018. Keras Documentation. Available at: https://keras.io/ [Accessed June 18, 2018].

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1), p.159–74.

LeCun, Y., Haffner, P., Bottou, L., Bengio, Y., 1999. Object Recognition with Gradient-Based Learning. In Springer, Berlin, Heidelberg, pp. 319–345.

Li, Y., Liu, F., 2016. Whiteout: Gaussian Adaptive Noise Regularization in Deep Neural Networks. Available at: http://arxiv.org/abs/1612.01490 [Accessed June 11, 2018].

Nilsson, N.J., 1971. *Problem-solving methods in artificial intelligence*, New York, NY: McGraw-Hill.

Rajpurkar, P., Hannun, A.Y., Haghpanahi, M., Bourn, C., Ng, A.Y., 2017. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. Available at: http://arxiv.org/abs/1707.01836 [Accessed June 11, 2018].

Samuel, A.L., 1959. Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.*, 3(3), p.210–229.

Silva, I., Epstein, M., 2010. Estimating loudness growth from tone-burst evoked responses. *J.*

Acoust. Soc. Am., 127(6), p.3629–42.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., et al, 2017. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. Available at: http://arxiv.org/abs/1712.01815 [Accessed June 11, 2018].

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.*, 15, p.1929–1958.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Stat. Soc.*, 36(2), p.111–147.

Sutton, G., Lightfoot, G., 2013. *Guidance for Auditory Brainstem Response testing in babies*, Reading. Available at: https://www.thebsa.org.uk/wp-content/uploads/2014/08/NHSP_ABRneonate_2014.pdf.

Vidler, M., Parkert, D., 2004. Auditory brainstem response threshold estimation: subjective threshold estimation by experienced clinicians in a computer simulation of the clinical test. *Int. J. Audiol.*, 43(7), p.417–29.

Wolpert, D.H., Macready, W.G., 1997. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.*, 1(1), p.67–82.

Neural Network Model	Accuracy ± STD (%)	F_1 Score
Convolutional	82.0 ± 5.9	0.821
Convolutional LSTM	79.2 ± 8.7	0.795
Convolutional Bidirectional LSTM	70.2 ± 12.3	0.705
Multilayer Perceptron	63.2 ± 9.6	0.632
Recurrent Neural Network (LSTM)	37.6 ± 12.1	0.379

Table 1. The performance of different neural network models as assessed by stratified 10-fold cross-validation. The mean accuracy and micro-averaged F_1 value across all three classes is presented.

Fold	Accuracy (%)	F_1 Score
1	90.0	0.900
2	90.0	0.900
3	80.0	0.800
4	85.0	0.850
5	75.0	0.750
6	79.0	0.789
7	73.7	0.737
8	88.9	0.889
9	76.5	0.765
10	82.4	0.824
Mean*/Micro-Average	**82.0***	**0.821**

Table 2. 10-fold cross-validation evaluation metrics for the best performing model. The accuracy and micro-averaged F_1 score for each fold is presented. The mean accuracy and micro-averaged F_1 value across all 10 folds is presented along the bottom row.

	Sensitivity (recall)	Specificity	PPV (precision)	NPV	F_1 Score	ROC AUC
RA	1.000	0.931	0.867	1.000	0.929	0.960
INC	0.875	1.000	1.000	0.929	0.933	0.936
CR	0.923	0.966	0.923	0.966	0.923	0.948
Micro-Average	**0.929**	**0.964**	**0.929**	**0.964**	**0.929**	**0.946**

Table 3. Final model performance on the test set. The performance of the model across all classes is compared. The micro-averaged values across all three classes are presented along the bottom row.

Figure Legends

Figure 1. Replicated ABR waveforms across a range of stimulus levels used to determine the estimated hearing threshold. Wave V is labelled where present. The ABR waveforms were recorded using the Interacoustics Eclipse. Image used with permission from Interacoustics A/S.

Figure 2. Final model architecture. The model is a deep convolutional neural network.

Figure 3. Training and validation accuracy across epochs. The training (black line) and validation (grey line) accuracy both improve until the model converges at around 1,000 epochs.

Figure 4. Confusion matrix of test set classification predictions. The predictions of the final model are compared to their correct labels.

Figure 5. Receiver Operating Characteristic curves. The ROC curves are presented for each class: 'response absent' (A), 'inconclusive' (B), and 'clear response' (C).

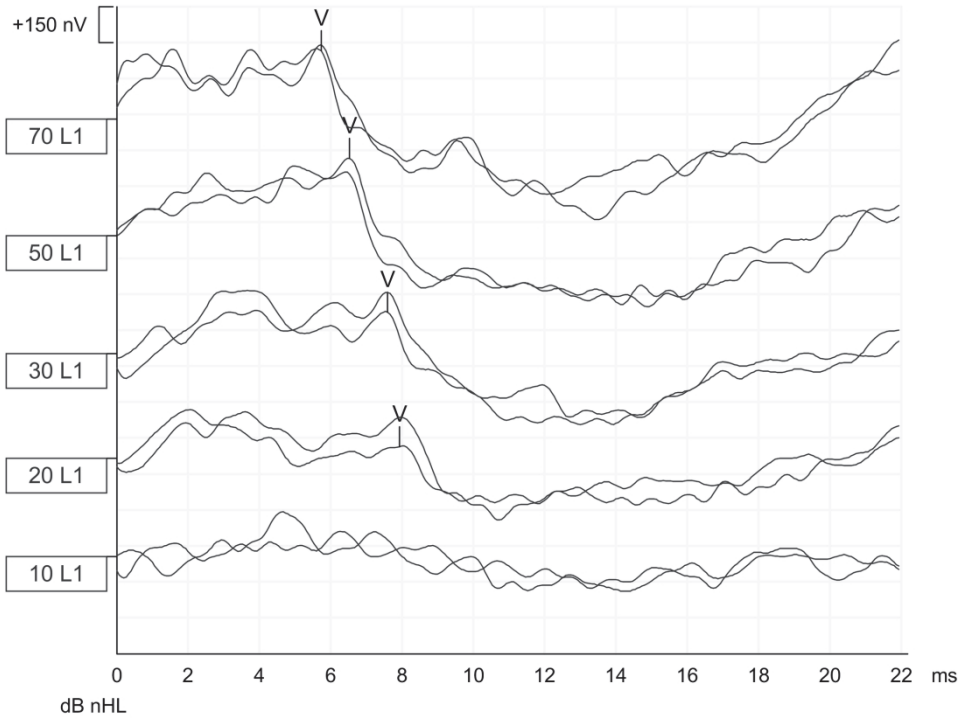
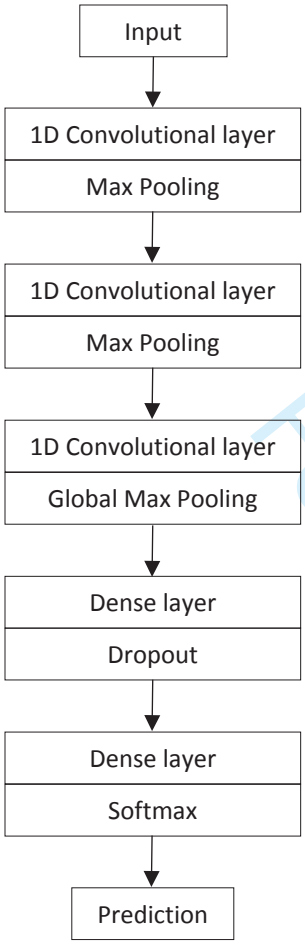


Figure 1. Replicated ABR waveforms across a range of stimulus levels used to determine the estimated hearing threshold. Wave V is labelled where present. The ABR waveforms were recorded using the Interacoustics Eclipse. Image used with permission from Interacoustics A/S.

99x76mm (600 x 600 DPI)



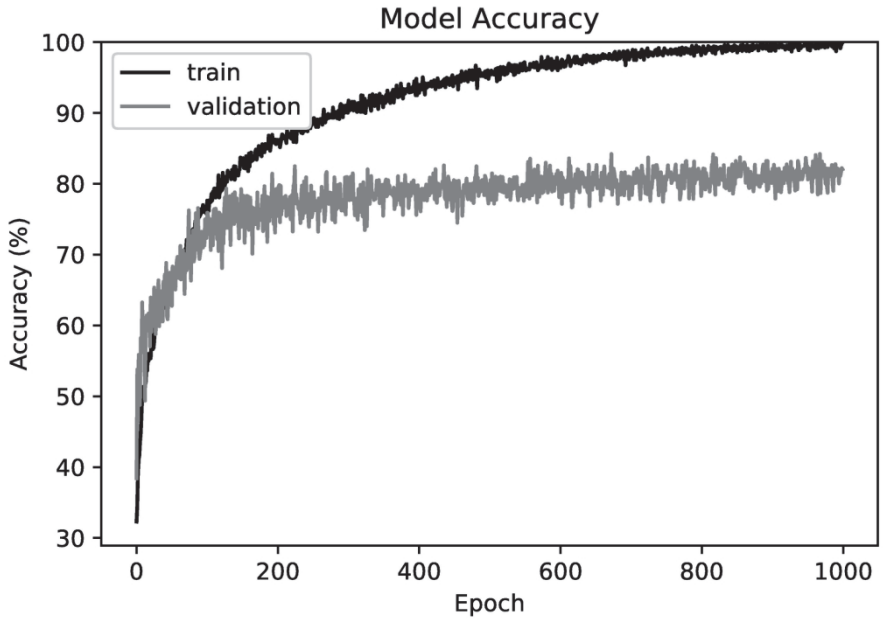


Figure 3. Training and validation accuracy across epochs. The training (black line) and validation (grey line) accuracy both improve until the model converges at around 1,000 epochs.

152x101mm (300 x 300 DPI)

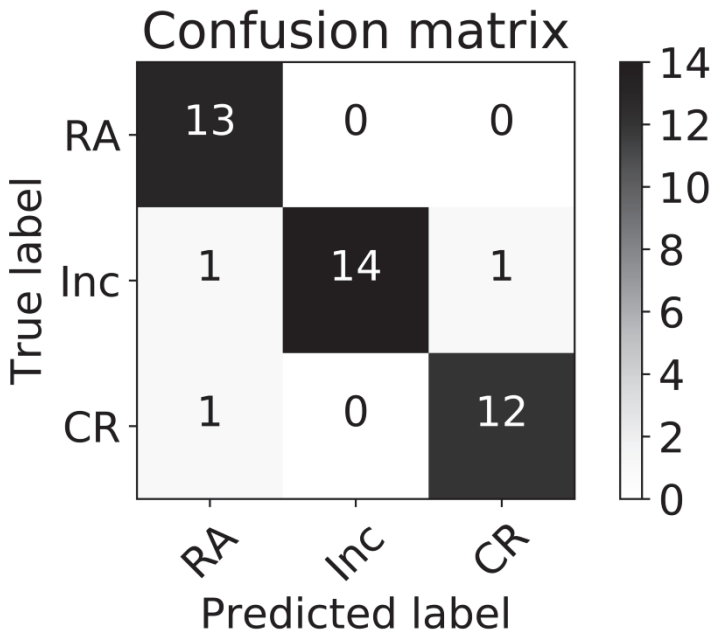


Figure 4. Confusion matrix of test set classification predictions. The predictions of the final model are compared to their correct labels.

152x101mm (600 x 600 DPI)

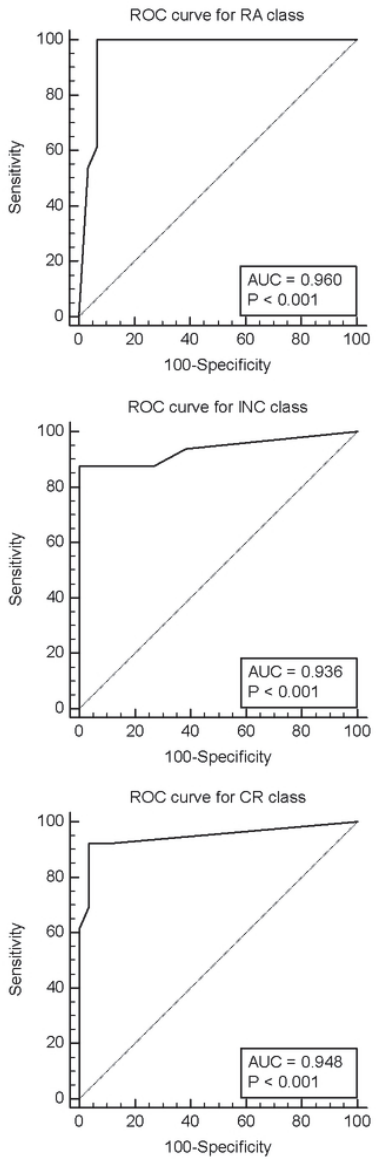


Figure 5. Receiver Operating Characteristic curves. The ROC curves are presented for each class: 'response absent' (A), 'inconclusive' (B), and 'clear response' (C).

15x43mm (600 x 600 DPI)