A Machine Learning Approach to Improve the Efficiency of the Frequency-Following Response Recording and Analysis

Master Thesis	Joint Master of Neuroscience
Student ID	21913639
Supervisors	José Valenzuela, Ph. D., Prof. Carles Escera

Brainlab - Cognitive Neuroscience Research Group
Institute of Neuroscience
Department of Clinical Psychology and Psychobiology
University of Barcelona
P. Vall d'Hebron 171, 08035 Barcelona, Spain

Table of contents

(word count +/-10%)

ABBREVIATION LIST

1D-CNN - One-Dimensional Convolution Neural Network

AABR – Automated Auditory Brainstem Response

ABR – Auditory Brainstem Response

Adam – Adaptive Moment Estimation

AEP – Auditory Evoked Potential

AI – Artificial Intelligence

ANN – Artificial Neural Network

AUC – Area Under the Curve

CNN – Convolutional Neural Network

CPU – Central Processing Unit

ConvNet – Convolutional Neural Network

DL – Deep Learning

EEG – Electroencephalography

ERP – Event-Related Potential

F0 – Fundamental Frequency

FFT – Fast Fourier Transform

FFR – Frequency Following Response

GPU – Graphics Processing Unit

IHS – Intelligent Hearing Systems

LLR – Long-Latency Responses

ML – Machine Learning

MLR – Middle-Latency Evoked Response

OAE – Otoacoustic Emissions

ReLU – Rectified Linear Unit Function

RF – Random Forest

ROC – Receiver Operating Characteristic

SNR – Signal-to-Noise Ratio

UNHS – Universal Newborn Hearing Screening

SCIENTIFIC BACKGROUND

The Frequency Following Response

The auditory system is responsible for the precision of sound encoding in the human brain, where it must quickly detect features such as onsets and offsets of acoustic signals, frequency, and amplitude to be able to make sense of the sounds around us. (Coffey et al., 2019)

An acoustic signal will be received by the sensory organ, travels on to the eardrum (tympanic membrane), where it reaches the malleus, incus, and stapes; the three bones of the middle ear (Fig. 1a)., which amplify the signal before it is transmitted to the organ of Corti in the cochlea (Fig. 1b). This structure harbors the hair cells, that perceive the sound vibrations and convert the mechanical signal into an electrical one (Fig. 1c). The information will be relayed further by the innervating nerve fibers to the central auditory nuclei in the brainstem until it finally reaches the cerebral cortex (Peterson et al., 2018) (Fig. 2a). The resulting neural response can be observed using electroencephalography (EEG), which detects voltage fluctuations that are time-locked to an event and thus called event-related potentials (ERPs) (Luck and Kappenman, 2011).

In addition, to observe ERPs, EEG is employed in various areas of neuroscience and biomedical engineering, such as seizure detection, sleep analysis, and brain-computer interfaces. It has great advantages over other methods, due to its non-invasiveness, its high temporal resolution as well as its relatively low financial costs (Craik et al., 2019) (Fig. 2b).

In auditory neuroscience, ERPs are usually triggered by acoustic stimuli such as clicks, syllables, speech, or music and are therefore referred to as auditory-evoked potentials (AEPs), which can be divided into three components. Because the precise timing of the sound is essential to localize where the stimulus is coming from, the auditory system is very sensitive to the timing of the acoustic signal. This temporal accuracy allows discrimination between auditory nerve and brainstem responses (ABRs), medium-latency evoked responses (MLRs), and long-latency responses (LLRs) recorded in the first 10 ms, between 10 and 60 ms, and between 60 and 200 ms after the onset of an auditory stimulus, respectively (Fig. 2c). Furthermore, we can distinguish between transient-evoked ABRs and steady-state frequency-following responses (FFRs), both of which are a signal of the ABR (Luck and Kappenman, 2011).

The comprehensive information that the FFR provides about how sound is processed in the brain is what sets it apart from other types of auditory-evoked neurophysiological responses.

Rather than only measuring the timing and amplitude, the FFR is able to recreate most of the characteristics of the incoming sound, reproducing much of the complexity of the eliciting stimulus (Kraus et al., 2017).

Moreover, the FFR acts as a window for the neural encoding of acoustic signals in the early stages of the auditory pathway (Xie et al., 2019) and can be used to investigate issues related to auditory processing impairment, autism, or neurodevelopmental speech and language disorders, as well as gender differences in auditory function (Coffey et al., 2019).

Various parameters in the time and frequency domain can be evaluated to define the characteristics of the FFR with sufficient accuracy. The time-domain includes the signal-to-noise ratio, the neural lag, the cross-correlation between the stimulus and the response as well as identification of pitch error and strength. In the frequency domain, the amplitude of the fundamental frequency (F0) and the corresponding harmonics can be retrieved (Ribas-Prats et al., 2019)

In most universal newborn hearing screening (UNHS) programs, a two-step screening process has been implemented. The faster and less expensive otoacoustic emissions (OAE) test is performed first, followed by an automated auditory brainstem response (AABR) test in neonates with certain risk factors or who do not pass initial screening (Patel and Feldman, 2011). The latter is using clicks and tone bursts to elicit a neural response, providing an objective characterization of hearing sensitivity. However, there are considerable limitations when it comes to revealing functional aspects of hearing, such as how environmental signals like speech are encoded by the brain (Madrid et al., 2021).

Since the FFR is sensitive to more complex auditory stimuli such as speech and music, detection of impairment of this specific ABR variant in newborns would allow early intervention and prevent many neonates from a lifelong struggle of disorders such as dyslexia (Ribas-Prats et al., 2019). Thus, a future FFR screening system that detects abnormal FFR signals early enough to suggest interventions is proposed.

Machine Learning

To create more independence from the need of trained professionals and to make the application of EEG more practical and less time-consuming, the scientific research community is working towards an automatic classification of neurophysiological signals using machine learning (ML) methods (Craik et al., 2019)

ML is a subfield of Artificial Intelligence (AI) (Fig. 3a) that deals with improving performance through experience, while AI addresses intelligence demonstrated by machines in general. ML can be further divided into three approaches of learning, namely supervised, unsupervised, and reinforcement learning (Fig. 3b).

A supervised learning algorithm is trained by receiving both the input data and the corresponding output data, which enables the computer to learn a function that allows it to predict the matching output with a certain accuracy. For example, the input data could be images showing a bus or pedestrians, which are then labeled "bus" or "pedestrian" respectively. Once the machine has learned the appropriate function for the training data, it can predict the label of a previously unseen image.

In unsupervised learning, the goal is to find patterns in the input data without prior training with labeled data. One of the most common examples is clustering, where the system is shown millions of images and recognizes a pattern that leads to a potentially useful cluster of similar images showing, for example, a cat (Russell and Norvig, 2021).

In reinforcement learning, the algorithm explores its environment and learns how to maximize its reward by determining which actions yield the best outcome (Sutton and Barto, 2018).

In the human nervous system, neurons are connected by their axons and dendrites through synapses. The strength of these synapses depends on the response to external stimuli and reflects the use of the connection, which is considered the brain's learning process. ANNs are inspired by this process, hence called “neural nets” and mimic this kind of learning (Aggarwal, 2018) (Fig. 3c).

Initially, ANNs failed to gain acceptance among neural classification methods due to practical problems such as long computation time and hardware bottlenecks. However, with recent developments in graphics processing units (GPUs) enabling powerful and cost-effective solutions, along with the availability of large datasets, ANNs and especially deep learning (DL) models are becoming increasingly popular.

Deep Learning

Deep Learning, a subfield of ML, is an ANN that consists of multiple processing layers and is capable of learning hierarchical representations of input data through sequential nonlinear transformations (Roy et al., 2019). This type of representation learning allows the ML algorithm to learn from raw data to automatically find the representations needed for a proper classification (LeCun et al., 2015). One can generally distinguish between three types of ANNs, depending on the nature of layers used: fully-linked, convolutional layers, or recurrent layers (Roy et al., 2019).

Imagine a supervised system fed with images to be classified into different categories, in which case the model will provide a result in the form of a vector of scores for each class. If the input image shows a car, we want the score for the "car" category to be the highest, which is unlikely without training. However, after such training, an objective function is computed, that measures the error between the true and observed class, and internal parameters called "weights" are adjusted accordingly to reduce the distance between observed and true classification scores (LeCun et al., 2015), which not only requires a large amount of training data but also depends on the nature of the input data. These weights are readjusted with each new batch of data fed to the algorithm. Reducing the prediction errors means moving toward the goal of achieving the highest possible accuracy for the classification task (Aggarwal, 2018), and ANNs have been shown to be capable of tackling this challenge.

One of the major advantages of DL systems over traditional ML algorithms is their ability to use raw or minimally pre-processed data as input, also known as end-to-end learning. The need to process the original signal to obtain related parameters or analogous information in a different and manipulable format is a limitation of classical ML algorithms and requires careful, often time-consuming engineering and considerable domain expertise to convert the raw data into suitable features that can be used as input for the subsequent classification task (LeCun, 2015) (Fig. 3d).

Moreover, features that are self-learned by the system often prove to be more effective than human-engineered features, as also highlighted by recent studies that were able to outperform their corresponding baseline(s) using raw EEG data as input to their DL models in terms of efficiency and accuracy (Roy et al., 2019).

HYPOTHESIS AND SPECIFIC AIMS

Considering that EEG signals are known to have a low signal-to-noise ratio (SNR), thousands of trials must be averaged to obtain a clear signal of the FFR (Yie et al., 2017). Acquiring the many trials needed to obtain a meaningful neural response from newborns is a time-consuming step, that poses a critical limitation (Hart and Jeng, 2021).

Neonatal EEG application is different from applying electrodes to cooperative adults. Typically, it is performed at the bedside, introducing artifacts due to the electromagnetically noisy background (Shany and Berger, 2011) and at some point, many newborns get irritated by the EEG cap or electrodes, and either a pacifier or breastfeeding needs to be used to calm them. The movements of the neonate's jaw lead to further artifacts, so an effective solution to this problem is to establish a method that requires only a few trials of stimulus presentation, also called sweeps, for an accurate neural signal (Xie et al., 2019). In the laboratory's most recent experiments, an average of 25 minutes was required to obtain the number of trials required for the FFR analysis, that is, 4000 sweeps (Arenillas-Alcón et al., 2021). Thus, improving experimental efficiency is one of the main motivations for using an ML-based approach (Xie et al., 2019).

Yi et al. (2019) demonstrated that vowel information can be reliably extracted from single-trial FFRs when ML principles are used, paving the way for further exploration. Furthermore, Craik et al. (2019) concluded in their review that classification using DL has been successfully applied to many EEG tasks, including ERPs, for both feature input and raw data. The latter has proven to be a great advantage, as self-learned features are often more informative and effective than human-engineered features (Roy et al., 2019).

Therefore, we hypothesized that a supervised end-to-end DL algorithm for classifying FFRs can achieve high levels of accuracy that will be the first methodological step to reducing the number of trials required, thereby shortening the time critical for screening in newborns.

Using the data previously collected in the laboratory, we aimed to determine, first, whether FFR signals elicited by two different stimuli could be classified by a DL model using raw data, though the limitation that we encountered will be discussed further in a later section.

Second, we wanted to assess whether the accuracy was high enough to be useful for subsequent development, as this model is part of the overall goal of predicting the neurodevelopmental level of participants at different ages.

The specific aim of this study was to develop a convolutional neural network (CNN) whose hyperparameters were tuned to achieve high accuracy in binary classification of brain responses to the stimuli /da/ and /oa/.

An explanation is given of how the model and its architecture were selected as the preferred DL method for the task at hand. Additionally, it is shown how parameters that can be changed by the researcher/developer, known as hyperparameters, are fine-tuned to achieve an accuracy above 95%.

MATERIALS AND METHODS

The Software

For the previous data recordings that the lab performed before the start of my internship, a SmartEP platform (Intelligent Hearing Systems (IHS), Miami, Fl, EEUU) was used, which generates binary files containing blocks of 50 trial averages. These .BLOCKS files had to be converted into editable files in a matrix format, such as .mat files, which was carried out by the lab technician using Matlab R2020b.

Python (version 3.7.13) was used to pre-process the data and implement the ML algorithm. Libraries and their versions are listed in (Fig. 4a). A Google Colab notebook served as the framework, which has the advantage of free GPU access (Bisong, 2019), speeding up the process compared to using a central processing unit (CPU) available on any computer.

The Data

The dataset comprises recordings from 86 neonates, of which 34 participants (17 females, aged 14-78 hours postpartum) were recorded using a two-vowel speech stimulus /oa/ by Arenillas-Alcòn et al. (2021). Additional data of 52 newborns (24 females, aged 14-125 hours postpartum) were provided by Ribas-Prats et al. (2019), who used the speech stimulus /da/ to elicit a neural response. All participants passed the UNHS test, conducted by the medical staff of the Hospital Sant Joan de Déu, followed by the EEG recording session performed by one of the researchers of the Brainlab. Once the newborn sleeps, the duration of the test session lasts about 25 minutes, including 5 minutes of preparation time for the recording.

For data acquisition, two electrodes were placed vertically on the forehead of the sleeping newborn and one reference electrode on each ear. The EEG was acquired at a sampling rate of 13,333 Hz, and an online bandpass filter was used in the range of 30 to 3000 Hz. In addition, any activity that exceeded ±30 µV was rejected online as an artifact (Ribas-Prats et al., 2019).

The data collection was arranged in 2 folders, one containing EEG recordings of the neural response to /oa/ stimuli (34 participants), the other to /da/ stimuli (52 participants). For each participant, there are 4 files containing information in the form of a matrix with dimensions 40x4096 (samples x time points), which sums up to a total of 13,760 samples (86 participants x 4 files per person x 40 samples). After inspecting the data frame, the matrix columns were truncated to a size of 3605 time points without losing critical information to prevent biased training of the machine learning model, which may have occurred due to the shorter length of the da recordings. To be used for training the algorithm, the data was labeled by adding a column with the value 0 or 1 for the used stimuli /oa/ or /da/ respectively (Fig. 4b).

The Model

CNNs are historically considered the most successful of all types of ANNs and are broadly used in computer vision, such as image recognition and object detection or localization, as well as in text processing (Nielson, 2015). The structure of a typical CNN architecture can be found in Fig. 4c.

The inspiration for this particular type of DL algorithm originated from Hubel and Wiesel's observations of the cat's visual cortex (Nielson, 2015), by introducing the theory that more complex receptive fields are formed by combining simpler receptive fields from cells at a lower level (Hubel and Wiesel, 1962).

This property, often referred to as sparse interaction, has been adopted by CNNs in that a given layer sees only a subset of the activations of the previous layer, achieved by using filters, also called kernels (Fig. 4d), that are smaller than the input image, which allows the NN to focus on spatially local features (Mattioli et al., 2022).

The second main characteristic of convolutional layers is parameter sharing, where the filter's connection weights are shared with other neurons (also called layer units or nodes) within the same layer to search for the same information across patches of the input, resulting in the model learning invariant representations of the data (Roy et al., 2019). This invariance, in turn, means that the predictions remain unchanged even after a transformation of the input variables, since, for example, with handwritten digits, a given number should always be predicted as such, regardless of its position in the image (Bishop and Nasrabadi, 2006).

One can imagine a filter that slides over all positions of a low-level input, detecting features in the receptive field of its size, and subsets of these so-called activations then serve as input to the next, higher layer. This process is usually repeated a few times to first detect primitive shapes such as edges or lines and then at higher levels to identify more complex structures such as loops (Fig. 5a). Mapping the activations from one layer to the next with a specific kernel is called a convolution operation, from which this neural network takes its name (Nielson, 2015).

Since CNNs are capable of extracting low-level features, the need for pre-processing is eliminated, making them very suitable for the approach of classifying EEG (Mattioli et al, 2022). In addition, Roy et al. (2019) found that more than 40% of studies that used DL methods for EEG analysis employed CNNs, and half of them used EEG data as raw or pre-processed time series data. We intended to work with raw data that did not require pre-processing before feeding it into the algorithm, but due to a limitation in the technical equipment used for acquisition, the data used in this study was already minimally pre-processed.

In the review by Kiranyaz et al. (2021) one-dimensional convolution neural networks (1D-CNNs) were found to be “the recent variants of conventional (2D-) CNNs”, especially for 1D signals. Moreover, they have compiled several articles proving that 1D-CNNs achieve state-of-the-art performance levels while keeping computational complexity to a minimum, making them easier to train.

RESULTS

Data Management

13,760 samples were split into training, test, and validation sets. 1,440 samples were retained as the validation set (10%), and 12,320 were used for training, which was further split into 3,080 samples for testing (23%) and 9,240 for training (67%) (Fig. 5b) using k-fold cross-validation (Fig. 5c).

In this case, K=4 was used to divide the training set into four equal parts, with a different part assigned as the test set and the rest as the training set in each round. This is done to minimize the risk of overfitting when evaluating the performance of the hyperparameters and then selecting the best parameter. After evaluating the hyperparameters, training and test datasets were pooled together and used as the training set for the model with all selected hyperparameters, and then the final performance of the model was evaluated by the validation set.

Model Training

Training the model worked best with an epoch size of 60, which indicates the number of training iterations above which accuracy stops increasing. The batch size, representative of the number of data windows exposed to the model before updating the model weights (Nielson, 2015), was set to the usual size of 100.

1D-CNN Architecture

The basic framework for the 1D-CNN was inspired by the most commonly used architectures for EEG signals. We selected some of the most important work in the field, entered the information into an Excel spreadsheet (Table 1), and chose the ones that best meet our needs, gradually changing the architecture until we achieved the best performance for our data, **98.12% classification accuracy**.

The model was structured into three blocks of layers, with the first and second parts being very similar in that both consist of two convolutional layers and a pooling layer followed by three batch normalization layers and a dropout layer (Fig. 6a). The last block consists of a flattening layer that converts the feature maps into a continuous linear vector, which is then passed to two dense layers, a batch normalization layer and a dropout layer. The number of filters used in the Convolution layers was 32, 64, 128 and 256 starting with the smallest number and increasing with each subsequent one. The associated filter sizes were 11, 11, 9, 7 respectively.

The running time of the model was 29 minutes and 30 seconds.

Dropout:

Many powerful neural networks suffer from overfitting (Fig. 6b), meaning that they perform well on training data but poorly on unseen (test) data. This happens when a large number of parameters learn very specific nuances of the training data that cannot be generalized, instead of detecting patterns in the data that can discriminate between different classes or clusters (Nielsen, 2015).

To tackle this problem, dropout has been proposed as the most common regularization technique for neural networks. It randomly modifies the network architecture by thinning the connections (Fig. 6c), which helps to avoid learning highly customized weights of learning data (Garbin et al., 2020).

More precisely, units of a neural network are taken out, that is, randomly selected nodes within the network are temporarily removed from the network along with their connections, by a certain percentage (Srivastava et al., 2014). Our network architecture includes three dropout layers, two of which are set to 0.2 and the last to 0.5, resulting in a strong improvement in reducing the difference in the loss function of the training and test data, thereby combating overfitting.

The loss function also called cost or error function, calculates the difference between predicted and actual values (Fig. 6d). The main goal of machine learning algorithms is to minimize this distance in order to make predictions with high accuracy (Nielson, 2015). Binary cross entropy, more specifically, the negative logarithm of the probability of the true class (Nielson, 2015), was used to calculate the loss.

Batch Normalization

Batch Normalization is considered one of the most successful architectural innovations in DL, as it enables higher learning rates and faster convergence of the network (Santurkar et al., 2018). It also promotes the independence of careful initialization of hyperparameters (initial weight values), which in turn leads to a reduction in training time (Garbin et al., 2020).

The most common belief is that its positive effects stem from stabilizing the distributions of the layer inputs by controlling the mean and variance of these distributions. This is challenged by the results of Santurkar et al. (2018), who claim that its effectiveness may be due to the property of batch normalization to significantly improve the loss and gradients in a model by smoothing the landscape of the corresponding optimization problem.

In our 1D-CNN we used batch normalization 7 times, as it was used after each convolutional, pooling, and dense layer, except for the final output layer, which showed a significant improvement shown in Fig. 7a.

Pooling

The purpose of pooling layers, of which we used two in our architecture, is dimensionality reduction. After the convolution operation has produced feature maps to find feature combinations, a merge is required to combine the detected features into one (LeCun et al., 2015). Different types of subsampling layers can be used, such as maximum pooling or average pooling. The latter proved to work best in our model, hence we used it twice with sizes of 5 and 7 (Fig. 7b).

Activation Function

An activation function is the output of a neuron given a set of inputs that is used to decide whether to activate the node. It is responsible for introducing nonlinearity into each neuron of the network, which is distinct from the graph of a linear classification system that outputs a plain line, and thus the new property of the network allows learning the complex patterns in the data. (Gustineli, 2022).

There are different activation functions used in different settings. In our model, we have chosen rectified linear unit function (ReLU) activation in all 4 convolutional layers and the first dense layer. For the final activation function, a sigmoid function was used as recommended for binary classification.

Adam

Adam or adaptive moment estimation is the most popular optimization algorithm (De Bardeci et al. 2021) that is used in many fields of science and engineering to find the optimal parameters of the model (Kingma and Ba, 2014). In our CNN, Adam optimization was used with a learning rate of 0.0001.

Performance Metrics

The assessment of the classification quality of the 1D-CNN was extended from accuracy alone to the use of a confusion matrix (Fig. 7c) and its derivatives, such as Precision and Recall, as well as AUC (Table 2).

DISCUSSION AND PERSPECTIVES

The results in this project are not only related to the final accuracy obtained by the CNN in its classification task but also to the system design itself and the dataset format. This is the reason why the description of how we obtained the final values of relevant hyperparameters, and also how we decided to use special techniques that allowed the system to achieve a high level of accuracy, can be found in the results section.

A 1D-CNN model with 4 convolutional layers, batch normalization, dropout, and two dense layers was implemented, which demonstrated that it is feasible to use a supervised end-to-end DL algorithm to classify FFRs while achieving high accuracy (98.12%). This is the first step in the development of a new recording and analysis system that will require fewer trials for classification in the future, shortening the time critical for newborn screening.

It is also the first step of many to achieve the goal of establishing a newborn screening system capable of identifying neonates who may later develop problems with reading and writing that could not previously be detected by the UNHS that is currently standard in European hospitals (Patel and Feldman, 2011).

While the experiment of finding a neural network that classifies the response elicited by two different stimuli has made considerable progress toward this goal, there are still two limitations that ought to be addressed in the future. At first, EEG recordings should be collected without any stimulus to provide the opportunity to train the model to discriminate between neural response and no response. This is necessary to detect newborns who are able to hear a stimulus but do not show the specific neural response of FFR, which has been proposed as a marker for neurodevelopmental issues such as dyslexia (Ribas-Prats et al., 2019).

The reason for the second limitation we encountered can be attributed to the deployment of the IHS system that was used to acquire the data, which by default averages 50 trials before the data can be retrieved. This issue has deprived us of the ability to assess whether our model is capable of classification based on single trial data, which could save even more crucial recording time. Thus, we propose to use a different setting in data recording or a distinct system that can be used to explore this option.

The next step in developing an early detection system after the "pure" classification between the presence and the absence of an FFR should be to develop a more sophisticated AI system in terms of the quality of sound processing. Rather than answering the question of whether there is a neural response, the model should predict how strong or weak the participant's FFR is. The evaluation of this information will be helpful for the third and final step to achieving the overall goal. A longitudinal study should be conducted in which the FFR of children is recorded at 0 months, 6 months, 12 months, and 24 months of age, along with the collection of additional information such as genetic data at all of these age groups.

The model is not only valuable for the goal of a screening application but also illustrates that it is indeed possible to use fewer trials, allowing future experiments to be performed much faster. In addition to the benefits of reducing the number of trials required, time is also saved by not being forced to use time-consuming feature engineering to analyze EEG data in terms of FFR, as the raw data can be used as it is to train the model without much pre-processing.

Although everything eventually worked out, we encountered some limitations, as mentioned above, and problems that we were not aware of before we started the experiment. One of the concerns that needed to be addressed was the unbalanced data set. Our dataset consists of data from 34 participants who were presented with the /oa/ stimulus and 52 who received the /da/ stimulus, thus the distribution of the data was not equal.

Model training required special care, as imbalanced data sets tend to bias classification in favor of the majority class, which means that the model is more likely to predict that an unseen data point belongs to the class that represents the majority of the total samples in the training data, and therefore frequently misclassifies examples from the minority class (Kumar et al., 2021).

Two possible solutions to overcome this problem have been suggested. Either the /da/ class is reduced to the same sample size as the /oa/ class, which results in the loss of data from which the algorithm can learn, and thus risking weaker performance. Or, the second proposed solution, involves combining a set of hyperparameters, such as batch normalization and dropout, which can be used to circumvent this constraint, however, in return, the performance measurement must be improved.

Using accuracy as the sole metric might lead to the assumption of good model performance, but when dealing with unbalanced data, appearances are often deceptive. As summarized by Luque et al. (2019), there are several performance metrics offered in the scientific literature to address this issue. While some are based on thresholds, probabilities, and ranks, the most widely used metrics are derived from the confusion matrix.

Roy et al. (2019) found that most studies used metrics such as accuracy, sensitivity, receiver operating characteristic (ROC) /area under the curve (AUC), and precision to evaluate the performance of the DL model. It was also mentioned that especially in binary problems, as in our case, it is important to make sure to use performance metrics that are robust to class imbalance, such as ROC/AUC. Therefore, we decided to use the confusion matrix and its derivatives, such as precision, recall, F1 score, AUC as well as false positives (misclassified /da/ stimulus responses) and false negatives (misclassified /oa/ responses) to provide an appropriate evaluation metric in addition to measuring the accuracy.

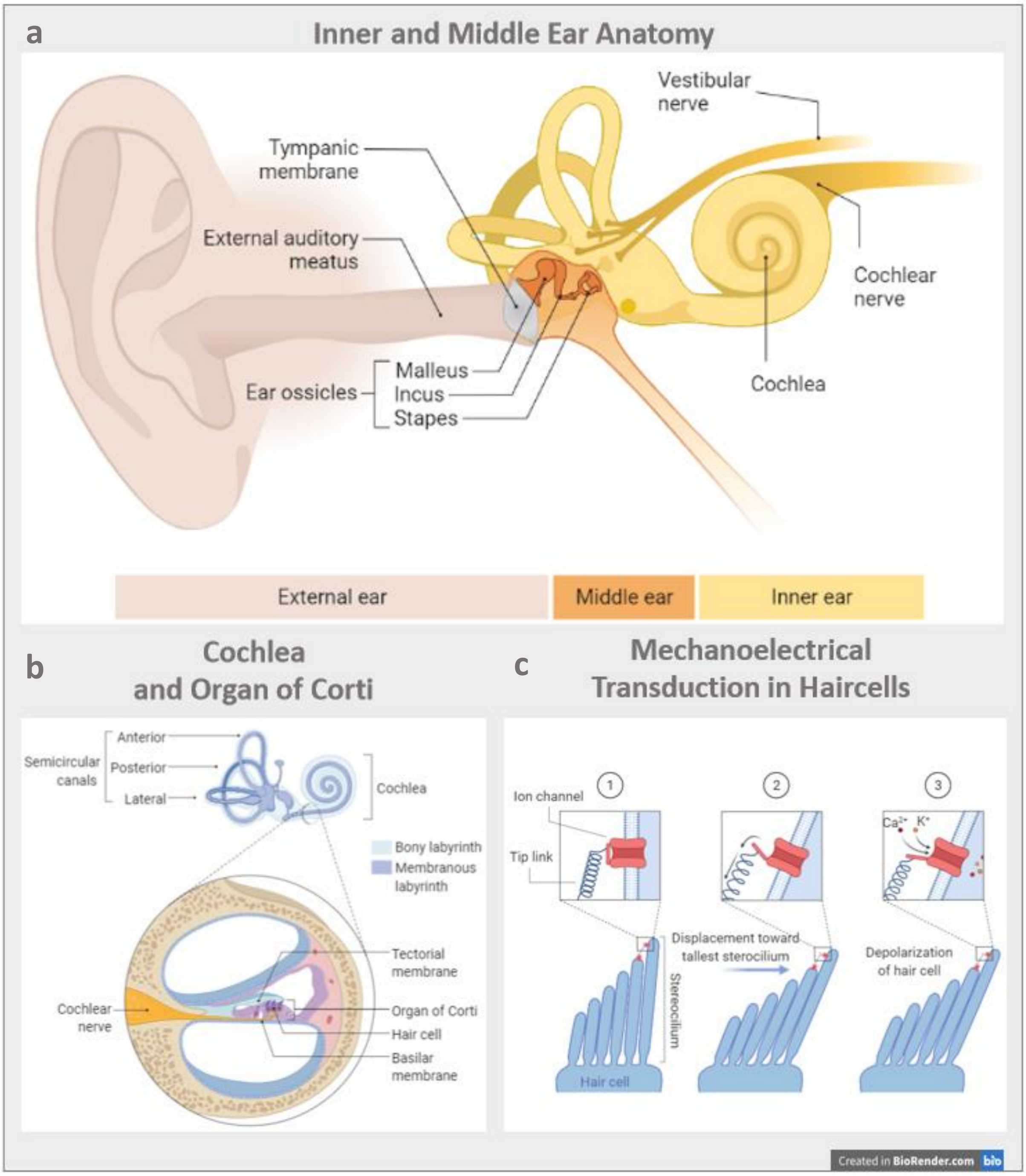


Fig. 1 | a. Anatomy of the inner and middle ear. An auditory stimulus travels along the external auditory meatus to the tympanic membrane, which transmits the vibration pressure to the three ossicles, which amplify the signal about 20-fold. It is further transmitted to the cochlea and causes the fluid inside to vibrate. | b. Cochlea and organ of Corti. The organ of Corti harbours the hair cells that are responsible for converting mechanical into electrical signals. | c. Mechanoelectrical transduction in hair cells. At the tip of each hair cell are stereocilia that move in the vibrating fluid and open the mechanically gated ion channels, causing an electrical signal.

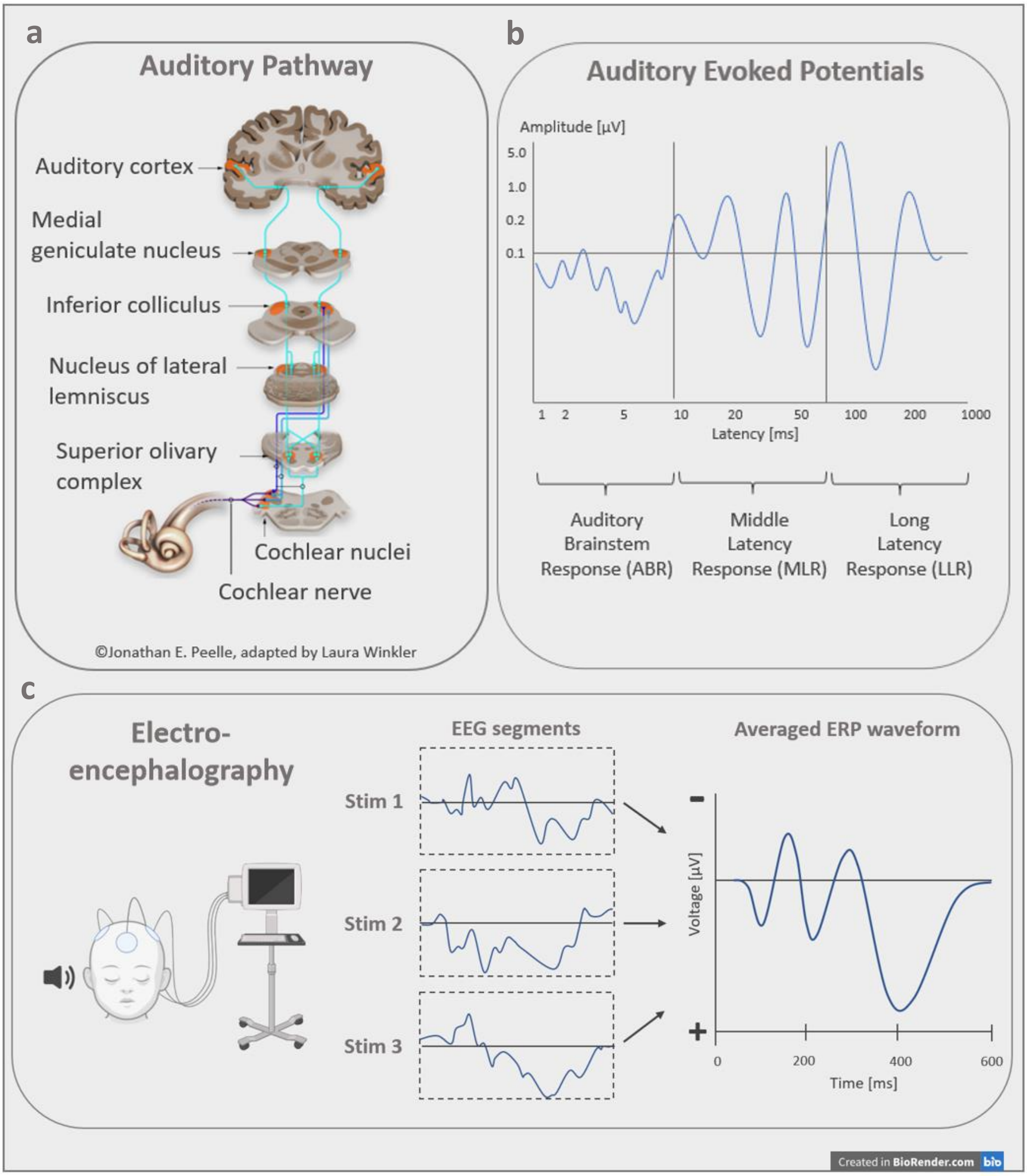


Fig. 2 | a. Auditory Pathway. The auditory signal travels along the cochlear nerve to the cochlear nuclei, on to the superior olive complex, the lateral lemniscus nucleus and the inferior colliculus. It is then passed on to the nucleus geniculatus medialis and from there it finally reaches the auditory cortex. | b. Auditory Evoked Potentials. A signal recorded while traveling from the ear to inferior colliculus has a short latency compared to its onset and is called ABR, whereas MLRs are found to have latencies respective to a signal in the medial geniculate nucleus. LLRs have been found to originate from a signal from the auditory cortex, but it must be mentioned that the boundary between MLR and LLR is not so clear. | c. Electroencephalography. After the EEG recording of many trials with one stimulus, all segments must be averaged to obtain a clear signal.

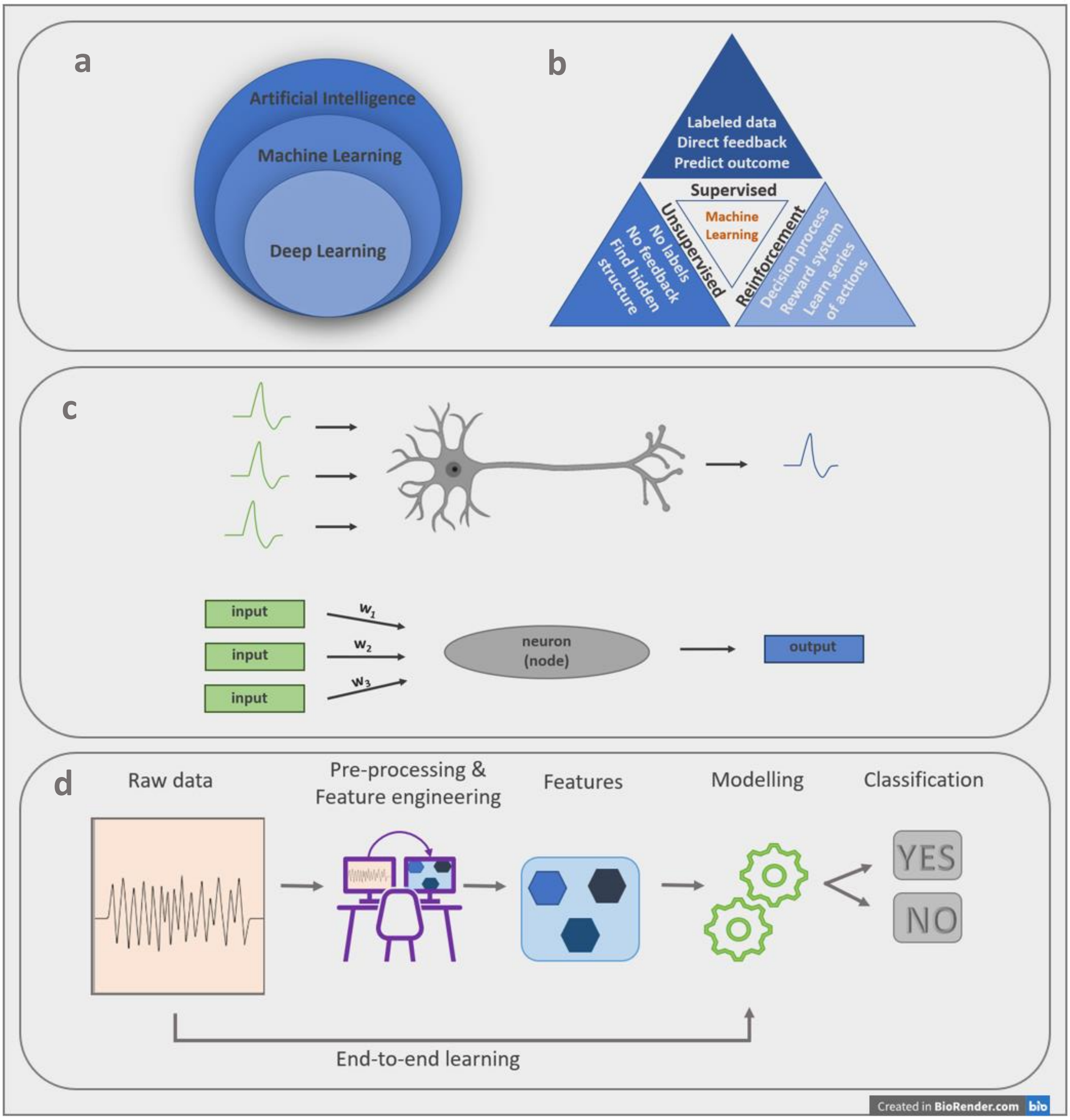
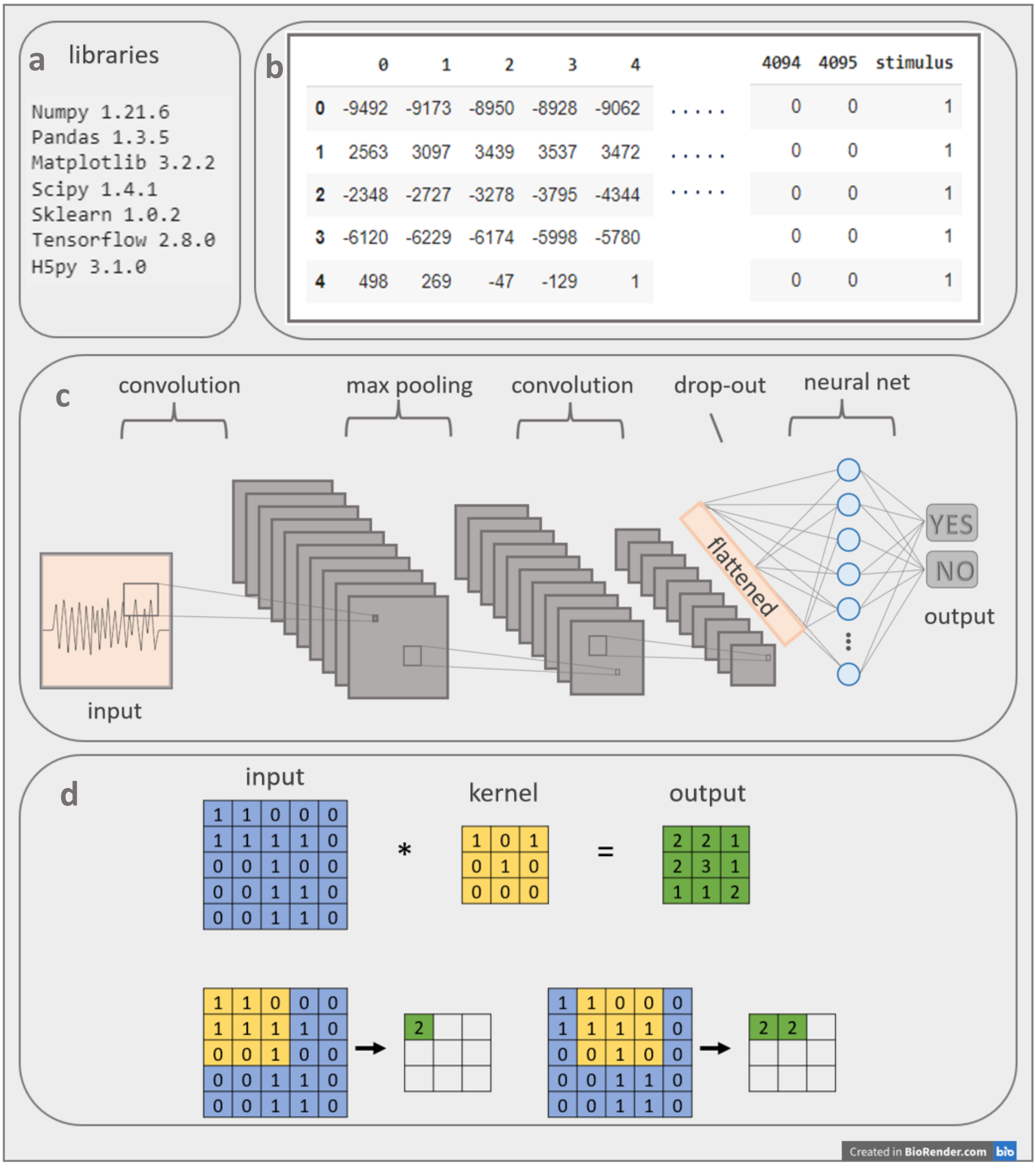


Fig. 3 | a. Relationship between Artificial Intelligence, Machine Learning and Deep Learning. AI describes the ability of machines to think, which means that a machine is able to make its own decisions without human intervention, while ML as a subfield refers to the development of systems that automatically learn from experience and improve without being explicitly programmed for the respective task. Deep Learning is even more specific and is the most advanced approach, using a neural network architecture with a large number of parameters and layers. | b. The Three Forms of Machine Learning. | c. A neuron of the human nervous system compared to an artificial neuron. The weights of the artificial NN represents the strength of the connection, based on the strength of a synapse of two biological neurons. | d. Pipeline of a supervised classification algorithm. The advantage of using an end-to-end learning model is apparent, as conventional ML methods require feature engineering as a preliminary step before modelling the algorithm.



```
Numpy 1.21.6
Pandas 1.3.5
Matplotlib 3.2.2
Scipy 1.4.1
Sklearn 1.0.2
Tensorflow 2.8.0
H5py 3.1.0
```

	0	1	2	3	4		4094	4095	stimulus
0	-9492	-9173	-8950	-8928	-9062	.....	0	0	1
1	2563	3097	3439	3537	3472	.....	0	0	1
2	-2348	-2727	-3278	-3795	-4344	.....	0	0	1
3	-6120	-6229	-6174	-5998	-5780		0	0	1
4	498	269	-47	-129	1		0	0	1

1	1	0	0	0
1	1	1	1	0
0	0	1	0	0
0	0	1	1	0
0	0	1	1	0

1	0	1
0	1	0
0	0	0

2	2	1
2	3	1
1	1	2

Fig. 4 | a. Python Libraries with their respective versions. | b. Part of the data frame using the Python library Pandas. The last column shows the number 1, which is the code label for the stimulus /da/. | c. Standard architecture of a CNN. The input can be either feature-based or raw data. Convolutional layers apply filters to input data in order to create feature maps. Pooling layers are used to reduce the dimensions of the feature maps and avoid overfitting. | d. Kernel. The filter slides over the input, multiplying each square with the number of the respective position of the input. Finally, the sum of all products results in the output of a square at the respective position of the newly created feature map.

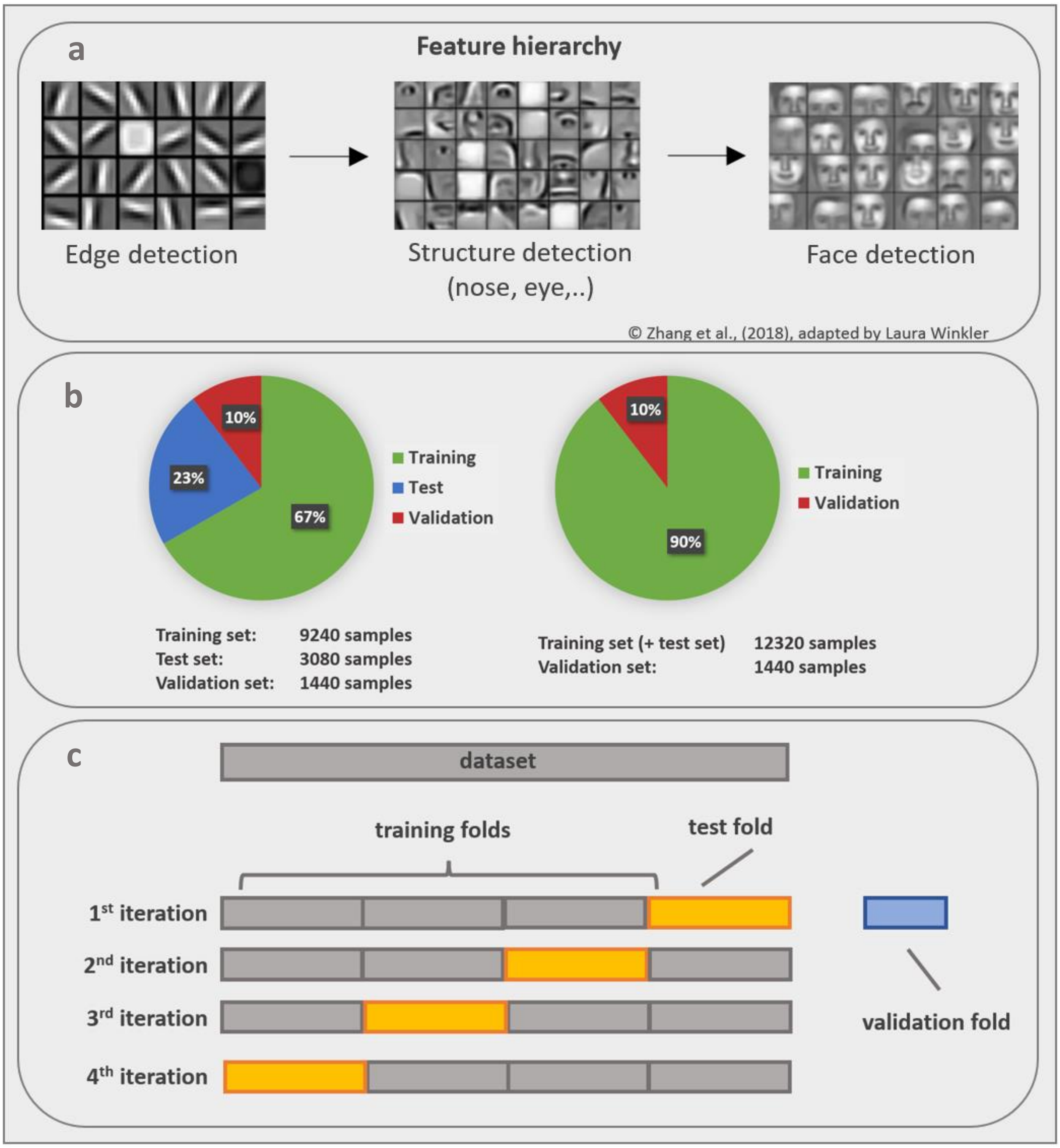


Fig. 5 | a. Feature hierarchy. Convolution close to the input data learns low-level features such as lines and edges, while convolutional layers deeper in the network architecture enable learning of higher-order features such as shapes or even specific objects such as noses or eyes, until finally faces are recognized. | b. Splitting the data set. For parameter tuning, the data set was split into a training set and a test set and then validated with the validation set. For the final training, the training and test set were combined and validated with the validation set. | c. K-fold cross-validation. A randomisation of the data is introduced, in which the training and test part of the data set is randomly split into a specific number of parts, here, a 4-fold cross-validation was used to avoid overfitting, with a different part being used as test subset at each iteration.

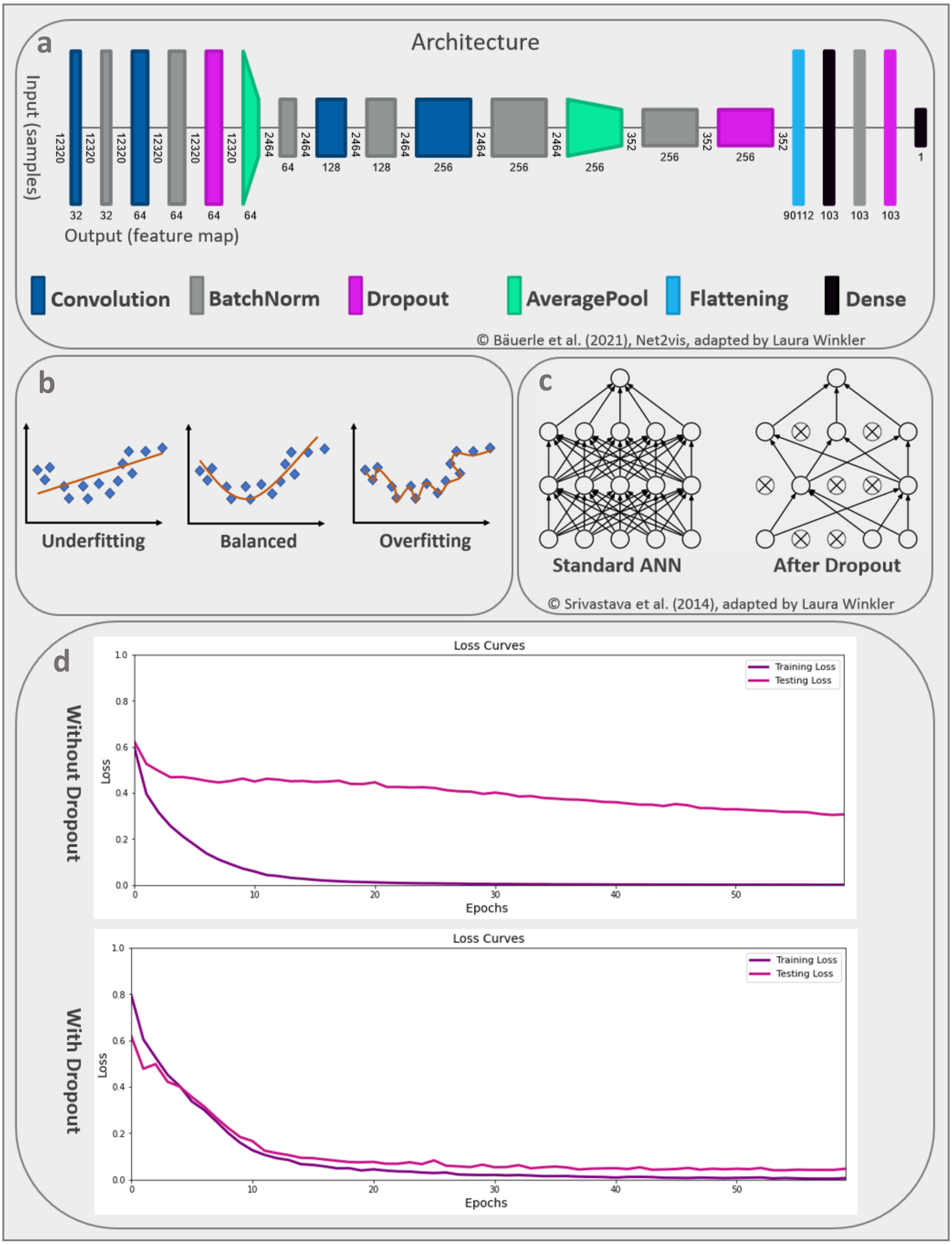


Fig. 6 | a. Architecture of the 1D-CNN model. | b. Underfitting, balanced, overfitting. A model that is either underfit or overfit to the training data will result in poor performance in classifying unseen data. | c. Dropout. This regularization technique is used to prevent overfitting by thinning the connections within the network. | d. Loss curves. The absence of dropout resulted in a difference of the loss between training and test set of 0.31, indicating overfitting, while a reduction of this difference of 0.04 can be observed when using this regularization technique.

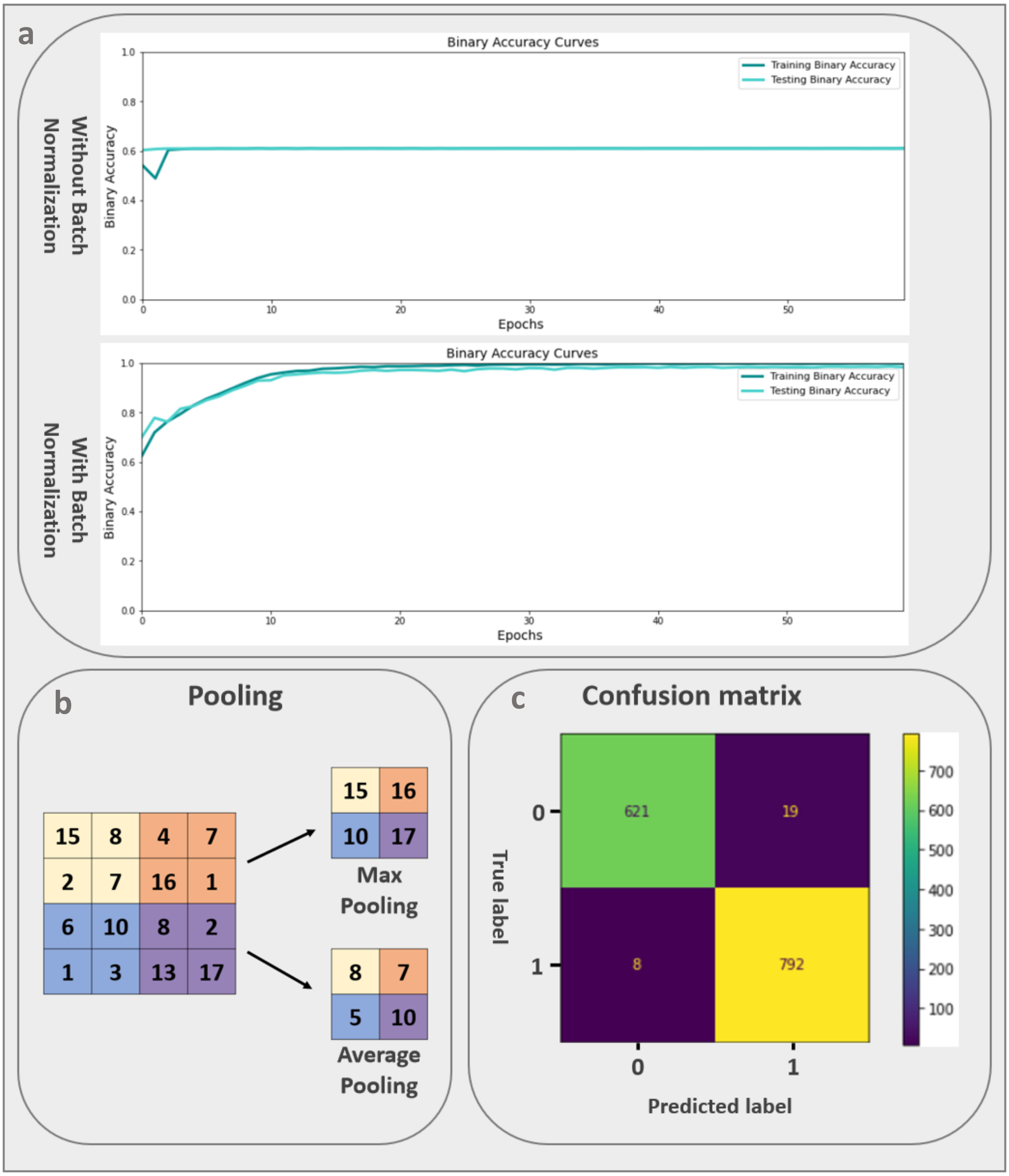
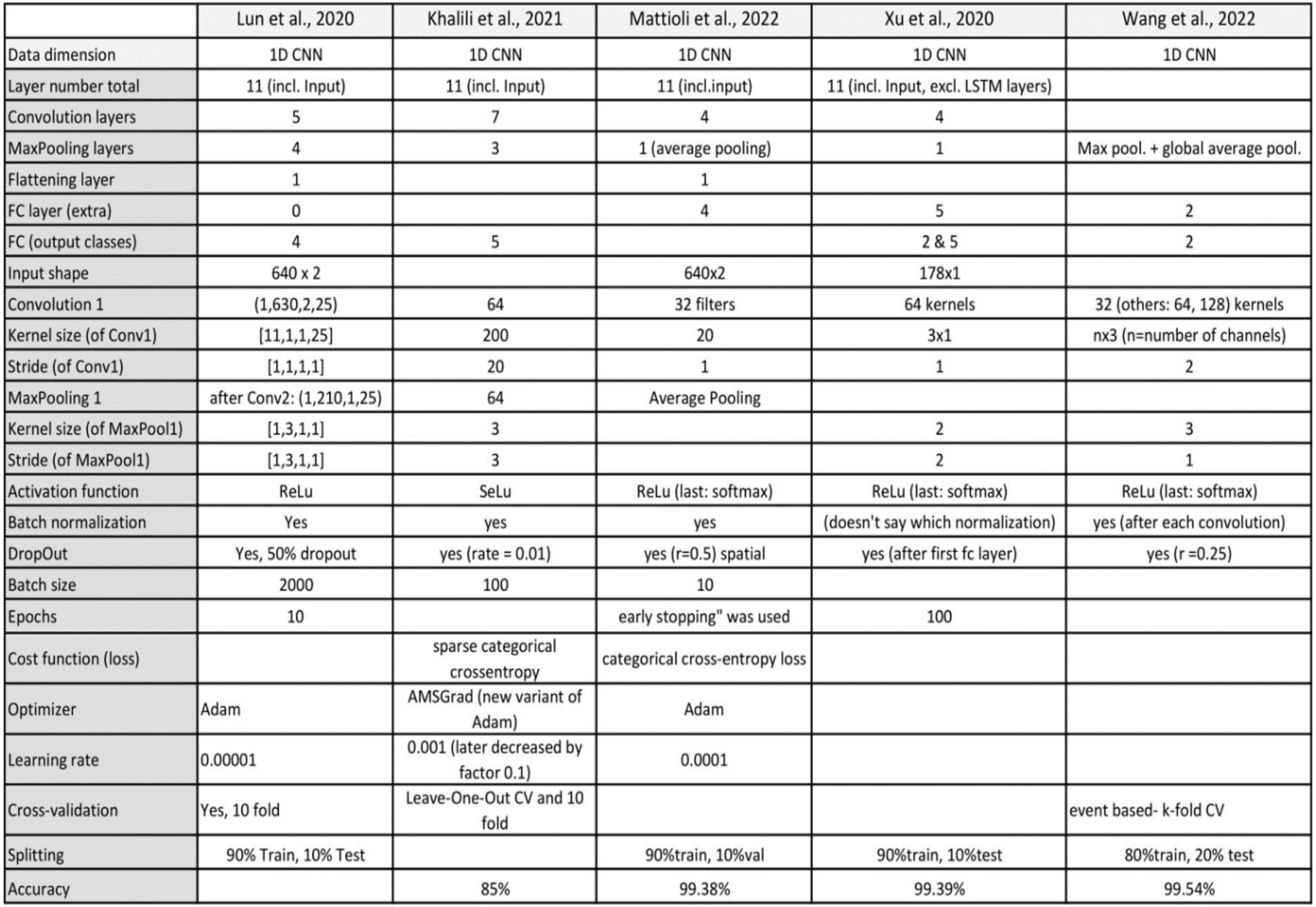
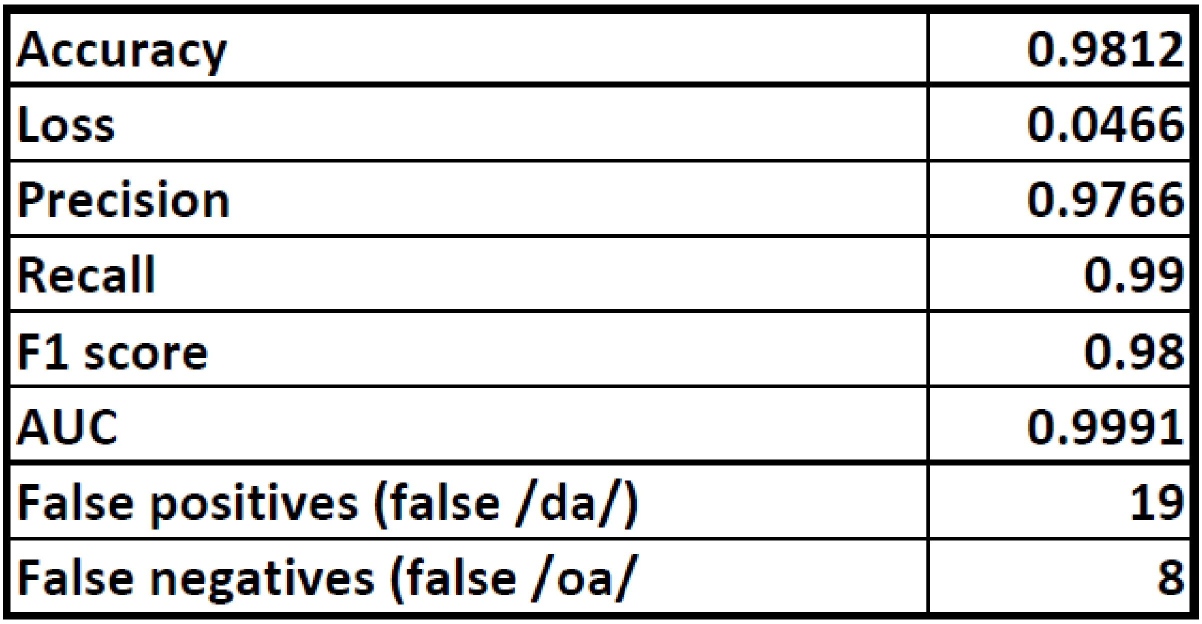


Fig 7. | a. Batch Normalization. Without batch normalization, the accuracy could not exceed 0.62, whereas when the proposed technique was applied, the accuracy improved significantly to above 0.95. | b. Pooling. While Max Pooling retrieves the highest number of a segment, Average Pooling, calculates the average of the section, which is then used as input for the next layer. | c. Confusion matrix. 621 of 640 /oa/ stimulus responses and 792 of 800 /da/ stimulus responses were correctly classified, reflecting an accuracy of 98.12%. 19 samples were misclassified as /da/stimulus responses, corresponding to the false positive values of the confusion matrix, while 8 false negatives represented the misclassified /da/stimulus responses.



	Lun et al., 2020	Khalili et al., 2021	Mattioli et al., 2022	Xu et al., 2020	Wang et al., 2022
Data dimension	1D CNN	1D CNN	1D CNN	1D CNN	1D CNN
Layer number total	11 (incl. Input)	11 (incl. Input)	11 (incl.input)	11 (incl. Input, excl. LSTM layers)	
Convolution layers	5	7	4	4	
MaxPooling layers	4	3	1 (average pooling)	1	Max pool. + global average pool.
Flattening layer	1		1		
FC layer (extra)	0		4	5	2
FC (output classes)	4	5		2 & 5	2
Input shape	640 x 2		640x2	178x1	
Convolution 1	(1,630,2,25)	64	32 filters	64 kernels	32 (others: 64, 128) kernels
Kernel size (of Conv1)	[11,1,1,25]	200	20	3x1	nx3 (n=number of channels)
Stride (of Conv1)	[1,1,1,1]	20	1	1	2
MaxPooling 1	after Conv2: (1,210,1,25)	64	Average Pooling		
Kernel size (of MaxPool1)	[1,3,1,1]	3		2	3
Stride (of MaxPool1)	[1,3,1,1]	3		2	1
Activation function	ReLu	SeLu	ReLu (last: softmax)	ReLu (last: softmax)	ReLu (last: softmax)
Batch normalization	Yes	yes	yes	(doesn't say which normalization)	yes (after each convolution)
DropOut	Yes, 50% dropout	yes (rate = 0.01)	yes (r=0.5) spatial	yes (after first fc layer)	yes (r =0.25)
Batch size	2000	100	10		
Epochs	10		early stopping" was used	100	
Cost function (loss)		sparse categorical crossentropy	categorical cross-entropy loss		
Optimizer	Adam	AMSGrad (new variant of Adam)	Adam		
Learning rate	0.00001	0.001 (later decreased by factor 0.1)	0.0001		
Cross-validation	Yes, 10 fold	Leave-One-Out CV and 10 fold			event based- k-fold CV
Splitting	90% Train, 10% Test		90%train, 10%val	90%train, 10%test	80%train, 20% test
Accuracy		85%	99.38%	99.39%	99.54%

Table 1. | Information on architecture and parameters of chosen studies using 1D-CNNs for EEG classification.



Accuracy	0.9812
Loss	0.0466
Precision	0.9766
Recall	0.99
F1 score	0.98
AUC	0.9991
False positives (false /da/)	19
False negatives (false /oa/	8

Table 2. | Performance metrics.

BIBLIOGRAPHY

Aggarwal CC (2018). Neural networks and deep learning. Springer, 10, 978-3.

Arenillas-Alcón S, Costa-Faidella J, Ribas-Prats T, Gómez-Roig MD, Escera C (2021). Neural encoding of voice pitch and formant structure at birth as revealed by frequency-following responses. Scientific reports, 11(1), 1-16.

Bäuerle A, Van Onzenoodt C, Ropinski, T (2021). Net2vis–a visual grammar for automatically generating publication-tailored cnn architecture visualizations. IEEE transactions on visualization and computer graphics, 27(6), 2980-2991.

Bishop CM, Nasrabadi NM (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer

Bisong E (2019). Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA.

Coffey EB, Nicol T, White-Schwoch T, Chandrasekaran B, Krizman J, Skoe E, Zatorre RJ, Kraus N (2019). Evolving perspectives on the sources of the frequency-following response. Nature communications, 10(1), 1-10.

Craik A, He Y, Contreras-Vidal JL (2019). Deep learning for electroencephalogram (EEG) classification tasks: a review. Journal of neural engineering, 16(3), 031001.

De Bardeci M, Ip CT, Olbrich S (2021). Deep learning applied to electroencephalogram data in mental disorders: A systematic review. Biological Psychology, 162, 108117.

Garbin C, Zhu X, Marques O. (2020). Dropout vs. batch normalization: an empirical study of their impact to deep learning. Multimed Tools Appl 79, 12777–12815

Gemein LA, Schirrmeister RT, Chrabąszcz P, Wilson D, Boedecker J, Schulze-Bonhage A, Hutter F, Ball T (2020). Machine-learning-based diagnostics of EEG pathology. NeuroImage, 220, 117021.

Gustineli, M. (2022). A survey on recently proposed activation functions for Deep Learning. arXiv preprint arXiv:2204.02921.

Hart B, Jeng FC (2021). A Demonstration of Machine Learning in Detecting Frequency Following Responses in American Neonates. Perceptual and Motor Skills. 2021;128(1):48-58.

HUBEL DH, & WIESEL TN (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology, 160(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Khalili E, Asl BM (2021). Automatic sleep stage classification using temporal convolutional neural network and new data augmentation technique from raw single-channel EEG. Computer Methods and Programs in Biomedicine, 204, 106063.

Kraus N, Anderson S, White-Schwoch T (2017). The frequency-following response: a window into human communication. The frequency-following response (pp. 1-15). Vol. 61 Springer, Cham.

LeCun Y, Bengio Y, Hinton G (2015). Deep learning. nature, 521(7553), 436-444.

Luck SJ, Kappenman ES (Eds.) (2011). The Oxford handbook of event-related potential components. Oxford university press.

Lun X, Yu Z, Chen T, Wang F, Hou Y (2020). A simplified CNN classification method for MI-EEG via the electrode pairs signals. Frontiers in Human Neuroscience, 338.

Luque A, Carrasco A, Martín A, de Las Heras A (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, 91, 216-231.

Kingma DP, Ba J (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kiranyaz S, Avci O, Abdeljaber O, Ince T, Gabbouj M, Inman DJ (2021). 1D convolutional neural networks and applications: A survey. Mechanical systems and signal processing, 151, 107398.

Kumar P, Bhatnagar R, Gaur K, Bhatnagar A (2021, March). Classification of Imbalanced Data: Review of Methods and Applications. In IOP Conference Series: Materials Science and Engineering (Vol. 1099, No. 1, p. 012077). IOP Publishing.

Madrid AM, Walker KA, Smith SB, Hood LJ, Prieve BA (2021). Relationships between click auditory brainstem response and speech frequency following response with development in infants born preterm. Hearing Research, 108277.

Mattioli F, Porcaro C, Baldassarre G (2022). A 1D CNN for high accuracy classification and transfer learning in motor imagery EEG-based brain-computer interface. Journal of Neural Engineering, 18(6), 066053.

Nielsen MA (2015). Neural networks and deep learning (Vol. 25). San Francisco, CA, USA: Determination press.

Patel H, Feldman M (2011). Universal newborn hearing screening. Paediatrics & child health, 16(5), 301–310.

Peterson DC, Reddy V, Hamel RN (2018). Neuroanatomy, auditory pathway. StatPearls Publishing.

Ribas-Prats T, Almeida L, Costa-Faidella J, Plana M, Corral M, Gómez-Roig M, Escera C (2019) The frequency-following response (FFR) to speech stimuli: A normative dataset in healthy newborns. Hearing Research 371:28-39.

Russell S, Norvig P (2021). Artificial intelligence: a modern approach. Pearson.

Roy Y, Banville H, Albuquerque I, Gramfort A, Falk T, Faubert J (2019) Deep learning-based electroencephalography analysis: a systematic review. Journal of Neural Engineering 16:051001.

Santurkar S, Tsipras D, Ilyas A, Madry A (2018). How does batch normalization help optimization?. Advances in neural information processing systems, 31.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

Shany E, Berger I (2011). Neonatal electroencephalography: review of a practical approach. Journal of child neurology, 26(3), 341-355.

Sutton RS, Barto, AG (2018). Reinforcement learning: An introduction. MIT press.

Wang X, Zhang G, Wang, Y, Yang , Liang Z, Cong F (2022). One-Dimensional Convolutional Neural Networks Combined with Channel Selection Strategy for Seizure Prediction Using Long-Term Intracranial EEG. International journal of neural systems, 32(02), 2150048.

Xie Z, Reetzke R, Chandrasekaran B (2019) Machine Learning Approaches to Analyze Speech-Evoked Neurophysiological Responses. Journal of Speech, Language, and Hearing Research 62:587-601.

Xu G, Ren T, Chen Y, Che W (2020). A one-dimensional cnn-lstm model for epileptic seizure recognition using eeg signal analysis. Frontiers in Neuroscience, 14, 1253.

Yi HG, Xie Z, Reetzke R, Dimakis AG, Chandrasekaran B (2017). Vowel decoding from single-trial speech-evoked electrophysiological responses: A feature-based machine learning approach. Brain and behavior, 7(6), e00665.

Zhang L, Wang S, Liu B (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.

ABSTRACT

The frequency-following response is an auditory brainstem response that can reproduce speech and sounds with high fidelity. Disruption of this signal leads to speech and language disorders and has thus been proposed as a potential biomarker for literacy impairment.

Considering the low signal-to-noise ratio for this type of electroencephalographic signal, many trials are required to average the responses to obtain a clear signal, which is not only time-consuming but also particularly problematic when working with neonates.

Over the past decade, several researchers have paved the way for the use of machine learning methods to improve the analysis of EEG signals. We, therefore, hypothesize that supervised ML algorithms are able to classify specific neural responses, called FFR, with a reduced number of trials without the loss of signal strength.

While traditional ML methods use predefined features that require time-consuming pre-processing of input data to be able to classify data, end-to-end learning algorithms have the advantage that features can be learned from raw or only minimally pre-processed input data.

To explore this property, a 1D Convolutional Neural Network was chosen based on its reputation of being a good candidate for time-series data, such as EEG data.

The dataset of 86 newborns who were presented two different stimuli /da/ and /oa/ to elicit an FFR was split into a training set, a test set, and a validation set to train and test the supervised model for the binary classification task and validate it on unseen data.

An accuracy of 98.12% was achieved for the binary classification of FFRs, demonstrating the feasibility of this approach. Other evaluation measures used showed similar levels of model performance, such as 99,91% AUC, 97.66% precision and 99% recall. With this finding, we are one step closer to the goal of establishing a newborn screening system capable of identifying newborns who may later develop difficulties with reading and writing that are not yet detected by the current universal newborn hearing screening in European hospitals.

Keywords: Frequency Following Response, Machine Learning, Deep Learning, Artificial Intelligence, Electroencephalography, Auditory Neuroscience