Diagnosis of Hearing Deficiency using EEG based AEP Signals: CWT and improved-VGG16 Pipeline

Md Nahidul Islam [Corresp., 1] , **Norizam Sulaiman** [1] , **Fahmid Al Farid** [2] , **Jia Uddin** [3] , **Salem A. Alyami** [4] , **Mamunur Rashid** [1] , **Anwar PP Abdul Majeed** [5, 6] , **Mohammad Ali Moni** [Corresp. 7]

[1] Faculty of Electrical and Electronics Engineering Technology, Universiti Malaysia Pahang, Pekan, Pahang, Malaysia

[2] Faculty of Computing and Informatics, Multimedia University, Cyberjaya 63000, Malaysia

[3] Technology Studies Department, Endicott College, Woosong university, Daejeon, South Korea

[4] Department of Mathematics and Statistics, Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia

[5] Innovative Manufacturing, Mechatronics and Sports Laboratory, Faculty of Manufacturing and Mechatronic Engineering Technology, Universiti Malaysia Pahang, Pekan, Pahang, Malaysia

[6] Centre for Software Development & Integrated Computing, Universiti Malaysia Pahang, Pekan, Pahang, Malaysia

[7] WHO Collaborating Centre on eHealth, UNSW Digital Health, University of New South Wales, NSW 2052, Australia

Corresponding Authors: Md Nahidul Islam, Mohammad Ali Moni
Email address: nahidul76.edu@gmail.com, m.moni@unsw.edu.au

Hearing deficiency is the world's most common sensation of impairment and impedes human communication and learning. Early and precise hearing diagnosis using electroencephalogram (EEG) is referred to as the optimum strategy to deal with this issue. Among a wide range of EEG control signals, the most relevant modality for hearing loss diagnosis is auditory evoked potential (AEP) which is produced in the brain's cortex area through an auditory stimulus. This study aims to develop a robust intelligent auditory sensation system utilizing a pre-train deep learning framework by analyzing and evaluating the functional reliability of the hearing based on the AEP response. First, the raw AEP data is transformed into time-frequency images through the wavelet transformation. Then, lower-level functionality is eliminated using a pre-trained network. Here, an improved-VGG16 architecture has been designed based on removing some convolutional layers and adding new layers in the fully connected block. Subsequently, the higher levels of the neural network architecture are fine-tuned using the labelled time-frequency images. Finally, the proposed method's performance has been validated by a reputed publicly available AEP dataset, recorded from sixteen subjects when they have heard specific auditory stimuli in the left or right ear. The proposed method outperforms the state-of-art studies by improving the classification accuracy to 96.87% (from 57.375%), which indicates that the proposed improved-VGG16 architecture can significantly deal with AEP response in early hearing loss diagnosis.

1 # Diagnosis of Hearing Deficiency using EEG based
2 # AEP Signals: CWT and improved-VGG16 Pipeline
3

5 Md Nahidul Islam[1], Norizam Sulaiman[1], Fahmid Al Farid[2], Jia Uddin[3], Salem A. Alyami[4],
6 Mamunur Rashid[1], Anwar P. P. Abdul Majeed[5,6], Mohammad Ali Moni[7]
7
8 [1] Faculty of Electrical and Electronics Engineering Technology, Universiti Malaysia Pahang,
9 Pekan, Pahang, 26600, Malaysia
10 [2] Faculty of Computing and Informatics, Multimedia University, Cyberjaya, 63000, MALAYSIA
11 [3] Technology Studies Department, Endicott College, Woosong University, Daejeon, South Korea
12 [4] Department of Mathematics and Statistics, Imam Mohammad Ibn Saud Islamic University,
13 Saudi Arabia
14 [5] Innovative Manufacturing, Mechatronics and Sports Laboratory, Faculty of Manufacturing and
15 Mechatronic Engineering Technology, Universiti Malaysia Pahang, Pekan, Pahang, Malaysia.
16 [6] Centre for Software Development & Integrated Computing, Universiti Malaysia Pahang, Pekan,
17 Pahang, Malaysia
18 [7] PhD, WHO Collaborating Centre on eHealth, UNSW Digital Health, University of New South
19 Wales, NSW 2052, Australia
20
21
22
23
24
25 Corresponding Author:
26 Md Nahidul Islam[1]
27 Mohammad Ali Moni[7]
28
29 [1]Faculty of Electrical & Electronics Engineering Technology, Universiti Malaysia Pahang,
30 26600 Pekan, Pahang, Malaysia
31 [7] PhD, WHO Collaborating Centre on eHealth, UNSW Digital Health, University of New South
32 Wales, NSW 2052, Australia
33
34 E-mail address: nahidul76.edu@gmail.com, m.moni@unsw.edu.au

36
37

Abstract

Hearing deficiency is the world's most common sensation of impairment and impedes human communication and learning. Early and precise hearing diagnosis using electroencephalogram (EEG) is referred to as the optimum strategy to deal with this issue. Among a wide range of EEG control signals, the most relevant modality for hearing loss diagnosis is auditory evoked potential (AEP) which is produced in the brain's cortex area through an auditory stimulus. This study aims to develop a robust intelligent auditory sensation system utilizing a pre-train deep learning framework by analyzing and evaluating the functional reliability of the hearing based on the AEP response. First, the raw AEP data is transformed into time-frequency images through the wavelet transformation. Then, lower-level functionality is eliminated using a pre-trained network. Here, an improved-VGG16 architecture has been designed based on removing some convolutional layers and adding new layers in the fully connected block. Subsequently, the higher levels of the neural network architecture are fine-tuned using the labelled time-frequency images. Finally, the proposed method's performance has been validated by a reputed publicly available AEP dataset, recorded from sixteen subjects when they have heard specific auditory stimuli in the left or right ear. The proposed method outperforms the state-of-art studies by improving the classification accuracy to 96.87% (from 57.375%), which indicates that the proposed improved-VGG16 architecture can significantly deal with AEP response in early hearing loss diagnosis.

Introduction

Hearing deficiency is the widespread form of human sensory disability; it is the partial or complete inability to listen to the ear's sound. The world health organization (WHO) reports that 466 million people were living with hearing loss in 2018, projected to exceed 630 million by 2030 and more than 900 million by 2050 ("Deafness and hearing loss"). An early and effective hearing screening test is essential for address the vast population concern. That helps to reduce the hearing deficiency by taking necessary steps at an appropriate time. Conventional listening tests and audiograms appear to be subjective assessments that significantly demand medical and health services. The audiogram reflects the hearing threshold across the speech frequency spectrum, usually between 125 and 8000 Hz. The traditional hearing impairment testing technique is very time-consuming, takes sufficient clinical time and expertise to interpret and maintain since it requires the person to respond directly. In the application of hearing aid, other issues, such as hearing loss's consequence (Holmes, Kitterick & Summerfield, 2017), the circumstances of the auditory stimulus (such as the background noise of the stimulus, locations of the stimulus (Das, Bertrand & Francart, 2018)(Das et al., 2016)), attention-altering techniques is still an open question.

Various hearing impairment testing techniques have been conducted to address these issues, and among them, EEG-based auditory evoked potentials (AEPs) are most widely used (Zhang et al., 2006)(Mahmud et al., 2019). Nowadays, the classification of AEP signal is most commonly used in many brain-computer interface (BCI) applications (Gao, Wang & Gao, 2014) and brain hearing issues (Sriraam, 2012). In fact, the AEP signal is widely used to recognize

79 hearing capability, assessment, and neurological hearing impairment identification. The AEP
80 signals are reflected by the brain's electrical activity changes in the body's sensory mechanisms in
81 response to the auditory stimulus. The diagnosis of hearing loss typically involves four main
82 stages: acquisition of data, data pre-processing, feature extraction and selection, and classification.
83 The feature extraction is traditionally conducted by analyzing the time-domain, frequency-domain,
84 and time-frequency domain techniques, which help to extract the information from the original
85 raw data. The extracted features are then used as an input to the machine learning or deep learning
86 models for training. However, traditional diagnosis methods have some drawbacks. For example,
87 traditional hearing loss approaches are often based on manual feature selection. As a consequence,
88 if the manually chosen features are ineffective for this task, the hearing loss recognition
89 performance will decrease considerably. Furthermore, Handcrafted features for different
90 classification tasks are task-specific, meaning that features that render predictions correctly are not
91 acceptable under certain conditions for other scenarios (Acir, Erkan & Bahtiyar, 2013)(Acir,
92 Özdamar & Güzeliş, 2006).

93   Although the researchers have employed a wide range of machine learning and deep learning
94 algorithms, recognizing the most effective classifier is still an open question. Among machine
95 learning-based classifiers, support vector machine (SVM) (Mahmud et al., 2019), k-nearest
96 neighbors (k-NN) (Thorpe & Dussard, 2018)(Rashid et al., 2021), artificial neural network (ANN)
97 (Mccullagh et al., 1996), linear discriminant analysis (LDA) (Grent-'t-Jong et al., 2021) Naïve
98 Bayesian (NB) (Shirzhiyan et al., 2019) are widely used in neurological response classification.
99 Nowadays, the convolutional neural networks (CNNs) are the most preferred approach in the
100 different classification tasks, particularly in image classification (Lecun, Bengio & Hinton, 2015).
101 In some recent studies, CNNs have shown promising performances in EEG signal classification:
102 in seizure detection (Ansari et al., 2019), depression detection (Liu et al., 2018), and sleep stage
103 classification (Ansari et al., 2018). Ciccarelli et al. (Ciccarelli et al., 2019) proposed a novel
104 architecture of the neural network and showed that their approach outperforms the linear methods
105 in decision windows of 10s. They have used eleven subjects in the experiment: with the wet EEG,
106 the decoding accuracy was improved from 66% to 81%, and with the dry EEG, the decoding
107 accuracy was improved from 59% to 87%. McKearney et al. (McKearney & MacKinnon, 2019)
108 used a deep neural network approach to classify paired auditory brainstem responses. They used
109 232 paired ABR waveforms (190 paired ABR waveforms for training the model and 42 paired
110 waveforms for performance evaluation) from eight normal hearing subjects and achieved 92.9%
111 testing accuracy. Although they achieved an excellent performance to identify the auditory
112 brainstem response, the testing set is too small, and more dataset is needed to test the model
113 performance. Mccullagh et al. (Mccullagh et al., 1996) reported a 73.7% accuracy using the
114 artificial neural network to classify 166 auditory brainstem responses (ABRs) with 2000
115 repetitions. Ibrahim et al. (Ibrahim, Ting & Moghavvemi, 2019) used multiple classification
116 techniques for detecting the hearing condition; the SVM algorithm outperforms the other
117 algorithms by achieving a classification accuracy of 90%. They used a nonlinear feature extraction
118 method to extract adequate information from the AEP signals. Dietl1 et al. (Dietl & Weiss, 2004)

119   evaluated an application to achieve detection of frequency-specific hearing loss where they used
120   the wavelet packet transform (WPT) as a feature extraction method and support vector machines
121   (SVM) classifier to transient evoked otoacoustic emissions (TEOAE). They achieved a maximum
122   of 74.7% accuracy with the testing dataset. Nonetheless, the overall accuracy is not favourable
123   enough to be utilized in real-life applications. Tang et al. (Tang & Lee, 2019) proposed a novel
124   hearing deficiency diagnosis method using three-level wavelet entropy, followed by MLP, trained
125   by hybrid Tabu search-Particle Swarm Optimization (TS-PSO). Their approach achieved 86.17%
126   testing accuracy; it still needs improvement for real-time applications. Sanjay et al. (Sanjay et al.,
127   2020) used machine learning approaches for human auditory threshold prediction. The absolute
128   threshold test (ATT) method was used for feature extraction from the auditory signals. The
129   extracted feature was then classified using multiple classification methods. Among all the
130   classification methods, a maximum of 93.94% accuracy was achieved with the SVM classifier.
131   Xue et al. (Xue et al., 2018) used participants' articulatory movements with or without hearing
132   impairment during nasal finals for hearing impairment diagnosis. Six different kinematic features:
133   standard deviation of velocity, minimum velocity, maximum velocity, mean velocity, duration,
134   displacement was used to extract the information from the hearing impairment (HI) patient and
135   normal hearing (NH) participants. The classification was conducted with a support vector machine,
136   radial basis function network, random forest, and C4.5. The maximum accuracy was 87.12% using
137   a random forest classifier via (displacement and duration feature). Zhang et al. (Zhang et al., 2006)
138   proposed an auditory brainstem response classification method. They used wavelet analysis for
139   feature extraction and Bayesian networks to classify the auditory responses. Discrete wavelets
140   transform (DWT) was used to extract the time-frequency information from the raw signals. A
141   maximum of 78.80% testing accuracy was achieved in their proposed approach; it needs more
142   improvement in testing accuracy.
143   The emphasis in our study is on a concise decision window. However, a concise window
144   contains less information and more difficult to achieve high performance but provide an effective
145   solution for early detection of hearing disorder. The short decision window is considered one of
146   the prerequisites to develop the real-life application, but limited studies have been carried out to
147   investigate this issue (Deckers et al., 2018). Moreover, selecting a short decision window makes
148   the system faster by reducing the computational complexity of the system. On the other hand, Deep
149   learning (DL) approaches can provide an effective solution because of their effective feature
150   learning capability to overcome the above limitations (Krizhevsky, Sutskever & Hinton,
151   2017)(Nossier et al., 2019)(Shao et al., 2019)(Bari et al., 2021)(Mahendra Kumar et al., 2021).
152   Deep learning models have several hidden layers that can explicitly learn hierarchical
153   representations. From model training, deep architectures can select discriminatory representations,
154   which are helpful for precise predictions according to the training data in subsequent classification
155   stages. Although the DL models have successful application in hearing loss diagnosis tasks, there
156   are still some issues with DL approaches. A few investigations (Ciccarelli et al., 2019)
157   (McKearney & MacKinnon, 2019) have been conducted using deep models with more than ten
158   hidden layers for hearing loss diagnosis. A large number of labelled data and computations

159 resources are typically required during the training model from scratch. In the proposed study, we
160 used the transfer learning (TL) method to address the challenges of training a deep model from
161 scratch. The TL method is used to expedite the deep learning model training phase and effectively
162 learns the hierarchical representations. The process is accomplished by using the pre-trained TL
163 method that has been pre-trained on vast datasets of natural images. The proposed pre-trained
164 model provides the lower-level weights for the target neural network, while the higher-level
165 weights are fine-tuned for the hearing deficiency diagnosis task. Consequently, the proposed TL
166 method offers a rational initialization for the target model and decreases the number of model's
167 parameters. In this manner, TL significantly enhances the performance of the training process.
168 Here, we summarized the main contribution of this paper.

169 • We have presented a hearing deficiency identification system based on deep CNN, where
170   a transfer learning strategy has been used to improve the training process. To fit the AEP
171   dataset in our model, we fine-tune the high-level parameters, consisting of unfreezing some
172   part of the pre-training model and re-training it. The lower-level parameters are transferred
173   from the previous trained deep architecture.

174 • In the proposed approach, we also changed some high-level parameters, reduced the
175   number of parameters and complexity of the TL architecture, which helps in improving the
176   performance of the VGG16 model for our dataset and reduces the computational time of
177   the training process.

178 • The experiment is conducted in a short decision window (1s and 2s), minimizing the impact
179   of additional features and reducing time consumption, which shows the proposed system
180   robustness and applicability in real-life application.
181

182 The rest of the manuscript is arranged as follows: a detailed data description, data pre-
183 processing, and the transformation process of CWT are implemented in the Materials and
184 Methodology section. A detailed description of the development of the proposed pre-trained model
185 and fine-tuning procedure for hearing deficiency diagnosis is also described in this section.
186 Experimental performance to determine the models' validation is described in the Result of the
187 Experiment and Analysis section. The Discussion section exhibits a discussion on the comparison
188 of the proposed model with related studies, along with the key advantages of our proposed method
189 over the previous studies. The Conclusion section represents the outcome of the present study.
190

Materials and Methodology

192 The aim of this study is to build an intelligent auditory sensation system for hearing loss diagnosis
193 with high performance. The overall procedure of the proposed hearing loss diagnosis method is
194 demonstrated in Figure 1. The proposed framework consists of few steps, including data collection,
195 pre-processing, time-frequency analysis, and building a pre-trained model with fine-tuning. We
196 have used a publicly available online dataset in the data collection phase instead of data collection
197 ourselves. We converted the raw signal into a time-frequency image using continuous wavelet

198 transform (CWT). Then, the proposed deep CNN (improved-VGG16) method is applied in the
199 time-frequency images for diagnosis the hearing loss. In the TL model, the pre-trained ImageNet
200 dataset has been used, and the size of the images is 224 * 224 pixels in RGB. The entire dataset
201 has been converted into a time-frequency image after data collection and resized in height-224 *
202 width-224 * depth-3. The VGG16 uses natural images which are different from the time-frequency
203 images of AEP. So, to fit the AEP dataset in the TL model, we replaced some VGG16 layers with
204 the new layers and then fine-tuned the improved VGG16 model.
205
206   Figure 1: The overall procedure of hearing deficiency diagnosis method.
207

Data Description
209 Experimental AEP datasets are provided by ExpORL, Dept. Neurosciences, KULeuven, and Dept.
210 Electrical Engineering (ESAT), KULeuven (Das, Francart & Bertrand, 2020). A 64-channel
211 BioSemi Active Two system was used for recording the AEP data, which was 8196 Hz sampling
212 rate. The entire data was collected from 16 normal-hearing subjects, and the trial was repeated 20
213 times from each subject. The recordings were conducted in a soundproof, electromagnetically
214 shielded space. The auditory stimuli were presented at 60 dBA by Etymotic ER3 insert earphones
215 and were low-pass filtered with a cut-off frequency of 4 kHz. As simulation software, APEX 3
216 was used (Francart, van Wieringen & Wouters, 2008). Three male Flemish speakers narrated four
217 Dutch stories as auditory stimulation ("Radio books for children"). Every story lasted 12 minutes
218 and was divided into two segments of 6 minutes each. Silent segments that lasted more than 500
219 milliseconds were shortened to 500 milliseconds. The stimuli were equal in root-mean-square
220 intensity and perceived as equally loud. The experiment was divided into eight sections, each
221 lasting six minutes. Subjects were presented with two parts of two storylines in each trial. The left
222 received one part, while the right ear received the other part. To prevent the lateralization bias
223 described by (Das et al., 2016), the attended ear was alternated over successive trials to ensure that
224 each ear received an equal volume of data. Each subject received stimuli in the same order, either
225 dichotically or after head-related transfer function (HRTF) filtering (simulating sound coming
226 from ±90°). As with the attended ear, the HRTF/dichotic condition was randomized and balanced
227 within and over subjects.
228

Data Preprocessing
230 The pre-processing of the AEP data is the first phase after data collection. In this study, the trials
231 were filtered with a high pass (0.5 Hz cut off) and downsampled from the sampling rate of 8192
232 Hz to 128 Hz. Here, we have investigated sixteen subjects, and each trial has been segmented into
233 the same length. The entire dataset has been segmented into short decision windows (1s and 2s)
234 and considered each decision window an observation. The straightforward reason to select the
235 concise decision windows is to reduce the computational complexity and make the system faster,
236 which will help detect the early hearing disorder. From each subject, 200 observations have been
237 picked, and finally, we achieved a total of 3200 observations. After data filtering and window

238 selection, the AEP data of subject-1, channel-1 in the time domain, is shown in Figure 2 when the
239 subject hears auditory stimulus through headphones defined as left and right labels.
240
241   Figure 2: AEP raw data plotting in 2s decision window: (A) hear auditory stimulus with the left
242   ear (B) hear auditory stimulus with the right ear.
243
244 **CWT for Time-Frequency Analysis**
245 CWT is a time-frequency feature extraction approach that offers multi-scale signal refinement by
246 scaling and translating operations. After the data pre-processing step, the segmented dataset
247 transforms from the time domain to the time-frequency domain using the CWT.
248 The CWT can automatically adapt the time-frequency signal analysis criteria and clearly explain
249 the signal frequency change with time (Yan, Gao & Chen, 2014). The CWT is widely used for
250 feature extraction and can be considered a mathematical tool for transforming time-series into a
251 different feature space. This study uses CWT as a feature extraction method that converts the raw
252 signal into 2-D time-frequency images from 1-D time-domain signals. An internal signal operation
253 and a series of wavelets are performed by the wavelet transforms. The mother wavelet is scaled
254 and translated to create the wavelet set, which is a family of wavelets $\psi(t)$, shown as
255

$$\psi_{S,\tau}(t) = \frac{1}{\sqrt{S}}\psi\left(\frac{t-\tau}{S}\right) \tag{1}$$

256
257 Here, S represents the scale parameter inversely related to frequency, and τ represents the
258 translation parameter.
259 The signal $x(t)$ can be achieved by a complex conjugate convolution operation, mathematically
260 defined as follows (Huang & Wang, 2018):
261

$$W(s,\tau) = \langle x(t), \psi_{S,\tau} \rangle = \frac{1}{\sqrt{s}} \int x(t)\psi^*\left(\frac{t-\tau}{S}\right)dt \tag{2}$$

262
263 Where $\Psi^*(\cdot)$ denotes the complex conjugate of the above function $\Psi(\cdot)$ and This operation
264 decomposes the signal $x(t)$ in a series of wavelet coefficients, in which the base function is the
265 wavelet family. In the equation, the s and τ are two types of parameters in the family wavelets.
266 The signal x(t) is transformed and projected to the time and scale dimensions of the family
267 wavelets.
268   In this study, we use wavelet basis functions (Mother Wavelets). The time-frequency
269 images are then used as the input of the proposed TL model. The transformation process of CWT
270 is shown in Figure 3.
271
272 Figure 3: The Transformation process from time-domain signal to time-frequency domain image.

273

274 Finally, we concatenate the 64 channels data in (M*M) for preparing an observation, where the
275 value of M is set to 8. So, each observation provides the time-frequency information of 64
276 channels. Figure 4 shows the time-frequency image of 64 channels.

277

278   Figure 4: The time-frequency image of 64 channels data.

279

280 **Hearing Deficiency Diagnosis using Deep TL.**
281 The proposed system presented a deep TL method based on improved-VGG16 architecture for
282 hearing loss diagnosis. The VGG16 uses natural images which are different from the time-
283 frequency images of AEP. The improvement consists of replacing some VGG16 layers with the
284 new layers and then fine-tuning the layers to fit the time-frequency AEP dataset in the model.

285

286 **Convolutional Neural Network Architecture**
287 LeCun (LeCun et al., 1998) proposed the convolutional neural networks (CNN), one of the best
288 pattern recognition methods. The locally trained filters are used in this system to extract the visual
289 features through the input image. CNN's internal layer structure consists of a convolution layer,
290 pooling layer, and fully connected layer. The complete procedure of CNN is shown in Figure. 5.

291

292   Figure 5: Typical Convolutional Neural Network Architecture.

293

294   •   Convolution Layer
295 The convolutional operations provide the more advanced feature representation. Several fixed-size
296 filters allow the complex functions to be used in the input image (Ravi et al., 2017). The same
297 weights and bias values are used in the whole image in each filter. This technique is called the
298 weight-sharing mechanism, and it makes it possible to represent the entire image with the same
299 characteristic. A neuron's local receptive field reflects the neuron's region in the previous layer.
300 This study uses the 'ReLU activation function (Nielsen). Let c × c is the size of the kernel or filter,
301 and i represent the time-frequency image. The weight and bias of the filter are denoted
302 by w and b, respectively. The output $O_{0,0}$ can be computed using Eq. (3), where f denotes the
303 activation function. This study used the ReLU activation function. In most of the classification
304 tasks, the ReLU activation function has demonstrated superior performance in terms of
305 accelerating convergence and mitigating the issue of vanishing gradients (Krizhevsky, Sutskever
306 & Hinton, 2017). The mathematical representation of the ReLU activation function can also be
307 seen in Eq. (4),

$$O_{0,0} = f\left(b + \sum_{t=0}^{c}\sum_{r=0}^{c} w_{t,r}i_{0+t,0+r}\right) \tag{3}$$

$$f(x) = \begin{cases} x & x > 0 \\ 0 & else \end{cases}. \tag{4}$$

308
309   • Pooling Layer
310   The pooling method is used in the feature maps, which have gone through convolution and
311   activation function. The pooling layer computes the local average or maximum value, reducing the
312   complexity and retaining the essential features, thus enhancing feature extraction performance.
313
314   • Fully connected layer
315   The convolutional and pooling layers alternately transfer the image features; after that, the fully
316   connected layer received the image feature as an input. One or more hidden layers may have in the
317   fully connected layer. By the data from the previous layer, each neuron multiplies the connection
318   weights and adds a bias value. Before transmission to the next layer, the measured value is passed
319   via the activation function. Eq. (5). displays neuronal calculations in this layer.
320

$$fc1 = f(b + \sum_{q=1}^{M} w_{1,q} * O_q) \tag{5}$$

321
322   Where f is the activation function, w is the weight vector, O is the input vector of the qth neuron,
323   and b is the bias value.
324
325   • SoftMax
326   The SoftMax activation function variates the logistic regression adapted to multiple classes and
327   used in the output layer for classification purposes. It can be determined by Eq. (6) (Sermanet et
328   al., 2013),
329

$$class_j = \frac{\exp(sf_j)}{\sum_q \exp(sf_q)} \tag{6}$$

330
331   **Proposed Pretrained Model Building and Fine-tuning**
332   In the convolutional neural network, the convolutional layers are used to extract the features from
333   the dataset in a different manner, whereas the fully connected layers are used to classify the
334   extracted features. The most forthright approach for enhancing the feature learning capability is to
335   increase the depth or width of the deep neural network. However, this can lead to two issues: the
336   first concern is that a deeper or wider model typically has more parameters, rendering the expanded
337   network more vulnerable to overfitting. The second concern is that it raises the use of computing
338   resources substantially.
339   To overcome these flowing issues and extract the AEP feature efficiently, the VGG16
340   network utilizes several parallel layers with different convolutional kernel sizes. It concatenates
341   the outputs at the end of the pre-trained network. In the proposed TL model, we replace some
342   layers of VGG16 with the new layers to fit the AEP dataset in the pre-trained network, which
343   enhances hearing loss identification performance. The replacement process consists of adding

344   some dense layers in the fully connected block of VGG16 architecture and adding the dropout
345   layers after every dense layer. A densely connect layer learns features from all the previous layer's
346   features. The dense layer performs a matrix-vector multiplication, and with the help of
347   backpropagation, the parameters can be trained and updated. The dense layer is used to change the
348   vector's dimensions and applies in other operations like rotation, scaling, and translation. Mele &
349   Altarelli (Mele & Altarelli, 1993) reported that on the CIFAT-10 dataset, the error rate 16.6%
350   when testing the dataset in a convolutional neural network. They improved the model's
351   performance with an error rate of 15.6% when the dropout layer was utilized in the last hidden
352   layer. We add the dropout layer after every dense layer in the fully connected block to reduce the
353   model complexity and prevent overfitting. The neuron is temporarily dropped with the probability
354   p at each iteration. Then, at every training step, the dropped-out neuron is resampled with the
355   probability p, and a dropped-out neuron will be active at the next step. Here, the hyperparameter p
356   is the dropout rate. Since the VGG16 uses the 'ImageNet' weight, which is trained with the natural
357   image, and the proposed time-frequency images are not similar, more layers need to be fine-tuned
358   where the weight is updated with the 'ImageNet' weight. This process helps to fit the time-
359   frequency images with the TL architecture. The proposed fine-tuning consists of unfreezing some
360   pre-trained network layers and re-train with the AEP dataset.
361   In the proposed approach, at first, we remove all the layers of VGG16 after the first 3x3
362   convolution layer of convolutional block 5, as shown in Figure 6, and replace the fully connected
363   block there. Then, we add multiple dense layers at the end of the VGG16 model, and after every
364   dense layer, we add the dropout layer. In the case of CNN, the convolutional layers extract the
365   feature from the dataset, whereas the fully connected layers try to classify the extracted features.
366   Consequently, adding more layers to the dense section can empower the network's robustness and
367   improve classification accuracy. So, despite using the two dense layers of the VGG16, here, we
368   add three new dense layers units of 1024, 512, and 288 in the fully connected block. Then, we add
369   a dropout layer after each dense layer, and the dropout value is set to 0.2, 0.4, and 0.6, respectively.
370   The reason behind adding the dropout layers is that the deep learning model reduces the
371   performance due to overfitting, and the dropout layers reduce the model complexity and prevent
372   overfitting. These techniques help in enhancing the performance in the hearing loss diagnosis. We
373   also remove the top layer and adding a SoftMax layer (output layer) based on the targeted class.
374   Based on the hyperparameters tuning technique, the proposed approach uses the 'Adam' optimizer
375   to adjust the network weight with the batch size 64, and the learning rate is set to 0.0001. The
376   parameters selection is made with the help of the 'Keras-Tuner' library. This library helps to select
377   the most optimal set of hyperparameters for our architecture. Hyperparameters are the variable that
378   governs the training process of the DL model and structure. There are two types of
379   hyperparameters: first, model hyperparameters that help in selecting the number and width of the
380   multiple hidden layers. Second, algorithm hyperparameters help to influence the speed and quality
381   of the learning algorithm. All the hyperparameters selected to build the proposed architecture are
382   based on ten different runs of the model. The following steps are used to train the model for hearing
383   loss identification, shown in Box 1.

384
385   Box 1: Training procedure of proposed TL architecture.
386
387   The detailed information of the parameter of the proposed TL architecture is shown in Table 1.
388   Here, C means the targeted class.
389
390   Table 1: Parameter of proposed TL architecture.
391
392   During the training process, all the layers before convolutional block four are frozen. The weights
393   are updated in the trainable layers, which helps in minimizing the errors between the predicted
394   labels and the actual labels. The complete architecture of the proposed TL has demonstrated in
395   Figure 6.
396   Figure 6: Transfer Learning Procedure of the Proposed Method.
397

Result of the Experiment and Analysis

399   This section represents the proposed hearing loss diagnosis method's performance based on CWT
400   and deep CNN architecture (improved-VGG16). First, we converted the time domain signal to
401   time-frequency domain images. Then, the images are resized into height-224 * width-224 * depth-
402   3, which is the suitable size of the proposed model. In this study, two different decision windows
403   were tested: 1s and 2s. This term refers to the quantity of data required to make a single left/right
404   decision. The practical reason behind selecting the shorter decision window is to detect the hearing
405   condition quickly. The entire dataset was randomly split into the training set and testing set. Here,
406   we used 70% dataset to train the architecture, and the rest of the dataset was used to test the model's
407   validation. This experiment has conducted with sixteen subjects where the subjects hear the
408   auditory track. Based on listening to the auditory track with the ear, the dataset has been divided
409   into two classes. The 'Class1' means the subject hears the auditory track with the left ear and the
410   'Class2' means the subject hears the auditory track with the right ear. With the (1s and 2s) decision
411   windows, we randomly selected 200 observations from each subject. A total of 2240 observations
412   has been used for training the model and 960 observations for testing the performance.
413
414   For 1s window length, the performance of the proposed approach for each subject in terms
415   of accuracy, precision, recall, f1-score and cohen's kappa of all subjects is demonstrated in Table
416   2.
417
418   Table 2: Performance of proposed model for 1s decision window.
419
420   Table 2 illustrates that in the case of subject-5, subject-7, and subject-16, our network achieves an
421   unprecedented performance of 100%. Except for six subjects (Subjects-3, 6, 9, 11, 13 and 14), all
422   subjects have achieved more than 90% accuracy. However, comparatively lower classification
423   accuracy has been noticed by Subjects-3 (86.67%), Subject-6 (83.33%), Subject-9 (76.67%),

424   Subject-11 (81.67%), Subject-13 (78.33%), and Subject-14 (76.67%). In the case of 1s decision
425   window length, the average classification accuracy is 91.56%, whereas the standard deviation is
426   8.91%. Besides classification accuracy, other performance evaluation techniques (such as
427   precision, recall, f1-score, and cohen kappa score) are also calculated to check the proposed
428   model's acuity. The average value of precision, recall, f1-score, and cohen kappa for sixteen
429   subjects are 90.74%, 93.63%, 91.92%, 82.71%, respectively, whereas standard deviations are
430   10.47%, 8.25%, 8.79%, 18.34%, respectively. Figure 7 shows the overall accuracy and loss curve
431   of the proposed TL method for the 1s decision window.
432
433   Figure 7: The overall accuracy and loss curve of the proposed TL method for 1s decision
434   window.
435
436   For 2s window length, the performance of the proposed architecture is illustrated in Table
437   3. In this case, a maximum of 100% accuracy has achieved for subject-6, subject-7, subject-10,
438   subject-16. Here, in the case of subject-16, we achieved 1.67% more accuracy compared to the 1s
439   time window analysis. However, the proposed architecture achieves an unprecedented
440   improvement (more than or equal to 90% for decision windows of 2s) in each subject. The lowest
441   accuracy of 90% has been obtained in subject-13.
442
443   Table 3: Performance of proposed model for 2s decision window.
444
445   With the 2s decision window, the average value of accuracy precision, recall, f1-score, and cohen
446   kappa for sixteen subjects are 96.87%, 96.49%, 97.57%, 97% and 93.73%, respectively. On the
447   other hand, the standard deviation of precision, recall, f1-score, and cohen kappa are 2.78%,
448   3.50%, 2.76%, 2.64% and 5.57%, respectively. Figure 8 shows the overall accuracy and loss curve
449   of the proposed TL method.
450
451   Figure 8: The overall accuracy and loss curve of the proposed TL method for 2s decision
452   window.
453
454   To illustrate the performance of the proposed TL model in depth, the confusion matrix of
455   all subjects has been given separately. A confusion matrix can be used to estimate the classification
456   accuracy of a model visually. Figure 9 represent the confusion matrix with 1s decision windows
457   analysis, whereas Figure 10 represent the confusion matrix with 2s decision window analysis. In
458   both figures, the letter A to P denotes the confusion matrix of subject-1 to subject-16, respectively.
459
460   Figure 9: Confusion matrix for 1s decision windows (A)subject1 (B)subject2 (C)subject3
461   (D)subject4 (E)subject5 (F)subject6 (G)subject7 (H)subject8 (I)subject9 (J)subject10
462   (K)subject11 (L)subject12 (M)subject13 (N)subject14 (O)subject15 (P)subject16
463

464   Figure 10: Confusion matrix for 2s decision windows (A)subject1 (B)subject2 (C)subject3
465   (D)subject4 (E)subject5 (F)subject6 (G)subject7 (H)subject8 (I)subject9 (J)subject10
466   (K)subject11 (L)subject12 (M)subject13 (N)subject14 (O)subject15 (P)subject16
467

468   The correct predictions are on the diagonal in the confusion matrix, while the incorrect predictions
469   are off the diagonal. For example, in the case of Figure 10(A) that denotes subject-1, a total of 59
470   observations (29 observations for class1, 30 observations for class2) have been recognized
471   accurately among 60 observations. In both decision windows, the total testing set for sixteen
472   subjects consists of 960 observations, in which 464 observations are in 'Class1', and 496
473   observations are in 'Class2'. For 1s decision windows, our network correctly detects 876
474   observations whilst 84 observations have been misclassified (shown in Figure 9). On the other
475   hand, for 2s decision windows, 930 observations have been accurately detected, whereas only 30
476   observations have been misclassified (shown in Figure 10). Therefore, 2s decision windows
477   provide a significant performance compared to the 1s decision windows.
478   Furthermore, to study the relationship between window length and detection performance,
479   this study includes a comparison. Figure 11 visualizes the average performance of two decision
480   windows over our network.
481

482   Figure 11: Hearing deficiency detection performance of the proposed TL architecture for two
483   different window lengths.
484

485

486   Figure 11 shows that the proposed TL network with a 2s decision window significantly improves
487   recognition accuracy compared to the 1s decision window analysis. The main goal of this study is
488   to enhance the performance for detecting the hearing condition with a concise decision window,
489   so that we can efficiently use this system in real-life application. For this purpose, first, we analyze
490   the 1s decision window and achieve 91.56% recognition accuracy; still not so high to apply this
491   system in real-life application. Furthermore, to enhance the performance of our proposed diagnosis
492   system, we move on to the 2s decision windows length, and this time we achieve a 5.31%
493   improvement in accuracy compared to the 1s decision window length. In the case of other
494   performance evaluation techniques such as precision, recall, F1 score and Cohen's kappa, we
495   achieve 5.74%, 3.94%, 5.08%, and 11.02%, improvement, respectively. The improvement
496   indicates the robustness and applicability of our proposed system.
497   Despite the impressive performance of the proposed system, in some cases, the
498   performance of our network is unsatisfactory. The possible reason for this poorer performance
499   compared to the other successful cases is that in EEG-based BCI application studies, a small SNR
500   and different noise sources are among the greatest challenges. Furthermore, Unwanted signals
501   contained in the main signal can be termed noise, artifact, or interference. Sometimes, the brain
502   may produce some unwanted noise due to the lack of the subject's proper attention or muscle
503   movement, affecting the detection results. In the experiment, we select concise decision windows
504   (1s and 2s), and working with a short window have many advantages but still very challenging

505 (Geirnaert, Francart & Bertrand, 2020). For these possible reasons, some subjects may provide a
506 lower accuracy compared to the other's subject (shown in Table2, and Table 3). Suppose in the 2s
507 decision windows length; if we avoid the two subjects that perform poorer than the other subjects
508 (shown in Table 3), we will achieve 97.62% recognition accuracy. However, the average training
509 and testing accuracy of sixteen subjects with 2s windows length is 100% and 96.87%, respectively,
510 after 100 epochs, whereas the standard deviation is 2.78%.

511   Furthermore, to study the robustness of the proposed method with a 2s decision window
512 (1s decision windows is not considered in the subsequent analysis), the performance of the
513 proposed model has been compared with other widely used TL architectures. Six popular transfer
514 learning algorithms namely, InceptionResNetV2 (Längkvist, Karlsson & Loutfi, 2014), MobileNet
515 (Pan et al., 2020), ResNet50 (He et al., 2016), VGG16 (Simonyan & Zisserman, 2015), VGG19
516 (Simonyan & Zisserman, 2015), and Xception (Chollet, 2017) have employed to the time-
517 frequency image of AEP dataset for hearing loss diagnosis. The input size is the same (height-
518 224* width-224* depth-3) for all the TL architectures. Figure 12 illustrates the performance
519 comparison of six popular TL models with the proposed model. According to Figure 12, the
520 proposed model achieved higher accuracy compared to the other TL models.

521

522   Figure 12: The performance comparison with other pre-trained architectures.

523

524 We also reduced the model parameters of VGG16 which help in reducing the model complexity
525 and minimize the computational resources. The total number of all model parameters and
526 performance is represented in Table 4. Table 4 reported that the overall accuracy is less than 61%
527 in all the pre-trained networks, where the models used pre-trained 'ImageNet' weights for hearing
528 impairment identification.

529

530   Table 4: Performance comparison with six popular TL models.

531

532 In the proposed TL methods (Improved-VGG16), we reduced the total number of parameters of
533 VGG16 (134,268,738 to 113,429,666). Although we reduced the number of parameters, the testing
534 accuracy was still improved to 96.87% from 57.37%. The reason behind the higher accuracy of
535 the proposed model compared to the other TL models is the replacement of some VGG16's layers
536 with the new layers and fine-tune the higher higher-level parameters, which helps to fit the AEP
537 dataset in the pre-trained network. This replacement consists of adding some dense layers in the
538 fully connected block of VGG16 architecture and adding the dropout layers after every dense layer
539 (shown in Figure 6). In the fine-tuning block, the time-frequency images are updated with the
540 'ImageNet' weight. This technique helps to fit the dataset in the proposed TL architecture and
541 enhance the overall performance for the hearing loss diagnosis. This experiment is carried out in
542 python, where we used Google colab, Windows 10, Intel(R) Xeon(R) CPU @ 2.30GHz, Tesla
543 K80, and CUDA Version: 10.1.

544   .

Discussion

A hearing deficiency detection method based on CWT and improved-VGG16 is proposed in this paper and achieved significantly outperform performance with the shorter decision windows (2s) than the previous state-of-art studies. The proposed improved-VGG16 architecture achieved an average accuracy, precision, recall, f1-score, and Cohen kappa of 96.87%, 96.50%, 97.58%, 97.01%, and 93.74%, respectively.

From Figure 12, it is clear that our network achieved more than 35% significant improvement compared to the others TL algorithms. In this experiment, we also found a significant effect of the decision window length on the overall performance. We achieved the improvement in the 2s decision window: 5.31% accuracy, 5.74% precision, 3.94% recall, 5.08% in F1 score, and 11.02% Cohen's kappa than the 1s decision window. The improvement is because the concise decision windows (1s) contain less information and sometimes provide unsatisfactory performance. However, this study aims to build an efficient network that can detect the hearing condition with a concise decision window so that we can able to achieve the decision quickly and can provide more effectiveness in real-life application.

Furthermore, a comparison of the proposed model with existing related studies is represented in Table 5. As seen in Table 5, Hallac et al. (Hallac et al., 2019) and Dass et al. (Dass, Holi & Soundararajan, 2016) utilized the convolutional neural network-based classification approach and achieved higher accuracy compared to the other related studies. Hallac et al. (Hallac et al., 2019) reported that with the raw AEP data and CNN, they achieved 94.1% accuracy. Dass et al. (Dass, Holi & Soundararajan, 2016) used both the time and frequency domain feature to extract the information from the raw AEP data. They used a feed-forward multilayer network to classify the AEP signal and achieved 90.74% testing accuracy. Both studies achieved a very encouraging performance but need more testing observations to validate the model's robustness.

Table 5: Performance comparison of related AEP studies.

In (Dietl & Weiss, 2004)(Mahmud et al., 2019) (Tan et al., 2013) (Li et al., 2019), the SVM classifier was used to classify the AEP dataset. Their approach achieved 78.80%, 85.71, 87%, and 78.7% accuracy, respectively. The obtained overall performance is not enough to apply the models in real-life application. Tang et al. (Tang & Lee, 2019) proposed a TS-PSO hybrid model to classify the two-class AEP dataset. They used Wavelet entropy as a feature extraction method and achieved 86.17% testing accuracy. Zhang et al. (Zhang et al., 2006) proposed a combination of wavelet analysis and Bayesian networks to classify auditory brainstem response (ABR) signals. For the wavelet analysis, they used the DWT method. Although they conducted an excellent analysis, the overall accuracy is reported 78.80%, which needs improvement.

The experimental outcomes demonstrated that the proposed architecture gain an impressive performance than the other related study for hearing deficiency diagnosis reported in the literature. Although the proposed approach outperforms state-of-art hearing deficiency detection methods, some difficulties are also faced during the experimental analysis. For example, to check the cross-

585 validation and prove the feasibility of our proposed network, a wide range of similar datasets is
586 needed. However, we did not find such dataset for further validation of the proposed method.
587 Another issue is the absence of clear speech envelopes in the dataset. In the previous research,
588 several types of EEG headsets were used to detect the hearing conditions, and these contain a
589 different number of electrodes (1-256). So, the number of electrodes and which electrodes are
590 required to achieve acceptable performance should be determined (Mirkovic et al.,
591 2015)(Montoya-Martínez, Bertrand & Francart, 2019)(Narayanan & Bertrand, 2018). In most of
592 the studies, the analysis is carried out with ordinary machine learning algorithms, and a few studies
593 are investigated with the deep learning approaches (Krizhevsky, Sutskever & Hinton,
594 2017)(Nossier et al., 2019)(Shao et al., 2019). However, most of the studies' testing accuracy is
595 not enough to use the model in real-time as well as real-life applications. A fast and more accurate
596 approach can be an efficient tool for future hearing devices and provide a great application in real-
597 life uses. Our study proposed the time-frequency distribution with a deep learning method and
598 achieved superior performance to other related approaches for hearing loss diagnosis reported in
599 the literature. The key advantages of our proposed method compared to previous studies are written
600 below:

601

602 • Instead of training the AEP dataset with the deep learning architecture from scratch, the
603 proposed study is conducted with a transfer learning strategy, which helps in faster training
604 and better accuracy.

605

606 • To fit our time-frequency AEP dataset with the pre-trained model weight, we fine-tuned
607 some higher-level parameters where the pre-trained weights are updated with the provided
608 dataset. This strategy helps in enhancing the overall performance for detecting hearing
609 deficiency.

610

611 • We compare the model's performance with the six popular TL methods, including VGG16
612 (Simonyan & Zisserman, 2015), VGG19 (Simonyan & Zisserman, 2015), MobileNet (Pan
613 et al., 2020), ResNet50 (He et al., 2016), InceptionResNetV2 (Längkvist, Karlsson &
614 Loutfi, 2014), and Xception (Chollet, 2017) algorithms where the proposed architecture is
615 superior for hearing deficiency diagnosis.

616

617 • We also changed some higher-level parameters (after the first layer of the convolutional
618 block 5, we remove all the layers and add the new fully connected layer shown in Figure
619 5). This approach also helps in reducing the VGG16 parameters and increasing the
620 performance of the proposed improved-VGG16 model.

621

622 • The proposed approach achieved the height classification accuracy of 96.87%, compared
623 to the previous studies (Ciccarelli et al., 2019) (McKearney & MacKinnon, 2019) (Ibrahim,
624 Ting & Moghavvemi, 2019) (Dietl & Weiss, 2004) (Tang & Lee, 2019) (Sanjay et al.,
625 2020) (Xue et al., 2018) (Zhang et al., 2006) (Tang & Lee, 2019) (Mahmud et al., 2019)

626 (Dietl & Weiss, 2004) (Zhang et al., 2006) (Tan et al., 2013) (Li et al., 2019) (Hallac et al.,
627 2019) (Dass, Holi & Soundararajan, 2016).
628

629 • The impact of different decision windows is also exhibited in the proposed study, whereas
630 our network provides a significant outcome with a concise decision window.

Conclusions

632 The proposed hearing loss diagnosis framework consists of two major steps: signal to image
633 transformation and building a hearing deficiency diagnosis system using deep CNN. In the
634 proposed study, the CWT is used to convert the raw signals to time-frequency images. Then, CNN-
635 based improved-VGG16 is used to classify the time-frequency images. This approach achieved
636 better outcomes with fewer trainable parameters, which help to reduce the training time of the
637 model. The applicability and effectiveness of the proposed method are verified by the publicly
638 available AEP dataset, and it achieved 96.87% testing accuracy with a concise decision window.
639 Moreover, this study will help to identify early hearing disorders efficiently. Because of the
640 unstable and subject-specific characteristics of the AEP signal, identification of the AEP signal is
641 challenging. Thus, to enhance the detection system's accuracy, other AEPs features need to be
642 investigated, and the use of more data variance and conditions can also be improved the outcome.
643

Acknowledgements

645 The authors would like to thank the Ministry of Higher Education for providing financial support
646 under Fundamental research grant No. FRGS/1/2018/TK04/UMP/02/3 (University reference
647 RDU190109) and Universiti Malaysia Pahang for laboratory facilities as well as additional
648 financial support under Internal Research grant RDU190109.

649

References

651

652 Acir N, Erkan Y, Bahtiyar YA. 2013. Auditory brainstem response classification for threshold
653 detection using estimated evoked potential data: Comparison with ensemble averaged data.
654 *Neural Computing and Applications* 22:859–867. DOI: 10.1007/s00521-011-0776-2.
655 Acir N, Özdamar Ö, Güzeliş C. 2006. Automatic classification of auditory brainstem responses
656 using SVM-based feature selection algorithm for threshold detection. *Engineering
657 Applications of Artificial Intelligence* 19:209–218. DOI: 10.1016/j.engappai.2005.08.004.
658 Ansari AH, Cherian PJ, Caicedo A, Naulaers G, De Vos M, Van Huffel S. 2019. Neonatal Seizure
659 Detection Using Deep Convolutional Neural Networks. *International Journal of Neural
660 Systems* 29. DOI: 10.1142/S0129065718500119.
661 Ansari AH, De Wel O, Lavanga M, Caicedo A, Dereymaeker A, Jansen K, Vervisch J, De Vos M,
662 Naulaers G, Van Huffel S. 2018. Quiet sleep detection in preterm infants using deep
663 convolutional neural networks. *Journal of Neural Engineering* 15. DOI: 10.1088/1741-
664 2552/aadc1f.
665 Bari BS, Islam N, Rashid M, Hasan J, Azraai M, Razman M, Musa RM. 2021. A real-time
666 approach of diagnosing rice leaf disease using deep learning-based faster R-CNN framework.
667 :1–27. DOI: 10.7717/peerj-cs.432.
668 Chollet F. 2017. Xception: Deep learning with depthwise separable convolutions. In: *Proceedings*

669 *- 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. Institute
670 of Electrical and Electronics Engineers Inc., 1800–1807. DOI: 10.1109/CVPR.2017.195.
671 Ciccarelli G, Nolan M, Perricone J, Calamia PT, Haro S, O'Sullivan J, Mesgarani N, Quatieri TF,
672 Smalt CJ. 2019. Comparison of Two-Talker Attention Decoding from EEG with Nonlinear
673 Neural Networks and Linear Methods. *Scientific Reports* 9:1–10. DOI: 10.1038/s41598-019-
674 47795-0.
675 Das N, Bertrand A, Francart T. 2018. EEG-based auditory attention detection: Boundary
676 conditions for background noise and speaker positions. *Journal of Neural Engineering* 15.
677 DOI: 10.1088/1741-2552/aae0a6.
678 Das N, Biesmans W, Bertrand A, Francart T. 2016. The effect of head-related filtering and ear-
679 specific decoding bias on auditory attention detection. *Journal of Neural Engineering* 13.
680 DOI: 10.1088/1741-2560/13/5/056014.
681 Das N, Francart T, Bertrand A. 2020. Auditory Attention Detection Dataset KULeuven. DOI:
682 10.5281/ZENODO.3997352.
683 Dass S, Holi MS, Soundararajan K. 2016. Classification of brainstem auditory evoked potentials
684 using artificial neural network based on time and frequency domain features. *Journal of
685 Clinical Engineering* 41:72–82. DOI: 10.1097/JCE.0000000000000148.
686 Deafness and hearing loss. *Available at https://www.who.int/news-room/fact-
687 sheets/detail/deafness-and-hearing-loss* (accessed June 18, 2020).
688 Deckers L, Das N, Ansari AH, Bertrand A, Francart T. 2018. EEG-based detection of the attended
689 speaker and the locus of auditory attention with convolutional neural networks.
690 *bioRxiv*:475673. DOI: 10.1101/475673.
691 Dietl H, Weiss S. 2004. Detection of cochlear hearing loss applying wavelet packets and support
692 vector machines. In: *Conference Record - Asilomar Conference on Signals, Systems and
693 Computers*. 1575–1579. DOI: 10.1109/acssc.2004.1399421.
694 Francart T, van Wieringen A, Wouters J. 2008. APEX 3: a multi-purpose test platform for auditory
695 psychophysical experiments. *Journal of Neuroscience Methods* 172:283–293. DOI:
696 10.1016/j.jneumeth.2008.04.020.
697 Gao S, Wang Y, Gao X. 2014. Visual and Auditory Brain – Computer Interfaces. 61:1436–1447.
698 Geirnaert S, Francart T, Bertrand A. 2020. An Interpretable Performance Metric for Auditory
699 Attention Decoding Algorithms in a Context of Neuro-Steered Gain Control. *IEEE
700 Transactions on Neural Systems and Rehabilitation Engineering* 28:307–317. DOI:
701 10.1109/TNSRE.2019.2952724.
702 Grent-'t-Jong T, Gajwani R, Gross J, Gumley AI, Krishnadas R, Lawrie SM, Schwannauer M,
703 Schultze-Lutter F, Uhlhaas PJ. 2021. 40-Hz Auditory Steady-State Responses Characterize
704 Circuit Dysfunctions and Predict Clinical Outcomes in Clinical-High-Risk Participants: A
705 MEG Study. *Biological Psychiatry*. DOI: 10.1016/j.biopsych.2021.03.018.
706 Hallac RR, Lee J, Pressler M, Seaward JR, Kane AA. 2019. Identifying Ear Abnormality from 2D
707 Photographs Using Convolutional Neural Networks. *Scientific Reports* 9:1–6. DOI:
708 10.1038/s41598-019-54779-7.
709 He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. In: *Proceedings
710 of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.
711 IEEE Computer Society, 770–778. DOI: 10.1109/CVPR.2016.90.
712 Holmes E, Kitterick PT, Summerfield AQ. 2017. Peripheral hearing loss reduces the ability of
713 children to direct selective attention during multi-talker listening. *Hearing Research*
714 350:160–172. DOI: 10.1016/j.heares.2017.05.005.

715 Huang L, Wang J. 2018. Forecasting energy fluctuation model by wavelet decomposition and
716   stochastic recurrent wavelet neural network. *Neurocomputing* 309:70–82. DOI:
717   10.1016/j.neucom.2018.04.071.
718 Ibrahim IA, Ting HN, Moghavvemi M. 2019. Formulation of a Novel Classification Indices for
719   Classification of Human Hearing Abilities According to Cortical Auditory Event Potential
720   signals. *Arabian Journal for Science and Engineering* 44:7133–7147. DOI: 10.1007/s13369-
721   019-03835-5.
722 Krizhevsky A, Sutskever I, Hinton GE. 2017. ImageNet classification with deep convolutional
723   neural networks. *Communications of the ACM* 60:84–90. DOI: 10.1145/3065386.
724 Längkvist M, Karlsson L, Loutfi A. 2014. Inception-v4, Inception-ResNet and the Impact of
725   Residual Connections on Learning. *Pattern Recognition Letters* 42:11–24.
726 Lecun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521:436–444. DOI:
727   10.1038/nature14539.
728 LeCun Y, Bottou L, Bengio Y, Haffner P. 1998. Gradient-based learning applied to document
729   recognition. *Proceedings of the IEEE* 86:2278–2323. DOI: 10.1109/5.726791.
730 Li PZ, Huang L, Wang CD, Li C, Lai JH. 2019. Brain network analysis for auditory disease: A
731   twofold study. *Neurocomputing* 347:230–239. DOI: 10.1016/j.neucom.2019.04.013.
732 Liu N, Lu Z, Xu B, Liao Q. 2018. Learning a convolutional neural network for sleep stage
733   classification. In: *Proceedings - 2017 10th International Congress on Image and Signal
734   Processing, BioMedical Engineering and Informatics, CISP-BMEI 2017*. Institute of
735   Electrical and Electronics Engineers Inc., 1–6. DOI: 10.1109/CISP-BMEI.2017.8302226.
736 Mahendra Kumar JL, Rashid M, Musa RM, Mohd Razman MA, Sulaiman N, Jailani R, Abdul
737   Majeed APP. 2021. The classification of EEG-based winking signals: A transfer learning and
738   random forest pipeline. *PeerJ* 9. DOI: 10.7717/peerj.11182.
739 Mahmud MS, Yeasin M, Shen D, Arnott SR, Alain C, Bidelman GM. 2019. What brain
740   connectivity patterns from EEG tell us about hearing loss: A graph theoretic approach. In:
741   *ICECE 2018 - 10th International Conference on Electrical and Computer Engineering*.
742   Institute of Electrical and Electronics Engineers Inc., 205–208. DOI:
743   10.1109/ICECE.2018.8636698.
744 Mccullagh PJ, King G, Mcallister HG, Houston HG. 1996. Classification of brainstem auditory
745   evoked potentials using artificial neural networks. *Studies in Health Technology and
746   Informatics* 34:547–550. DOI: 10.3233/978-1-60750-878-6-547.
747 McKearney RM, MacKinnon RC. 2019. Objective auditory brainstem response classification
748   using machine learning. *International Journal of Audiology* 58:224–230. DOI:
749   10.1080/14992027.2018.1551633.
750 Mele B, Altarelli G. 1993. Dropout: A Simple Way to Prevent Neural Networks from Overfittin.
751   *Physics Letters B* 299:345–350. DOI: 10.1016/0370-2693(93)90272-J.
752 Mirkovic B, Debener S, Jaeger M, De Vos M. 2015. Decoding the attended speech stream with
753   multi-channel EEG: Implications for online, daily-life applications. *Journal of Neural
754   Engineering* 12. DOI: 10.1088/1741-2560/12/4/046007.
755 Montoya-Martínez J, Bertrand A, Francart T. 2019. Optimal number and placement of EEG
756   electrodes for measurement of neural tracking of speech Dynamical Systems, Signal
757   Processing and Data Analytics Oto-Rhino-Laryngology Optimal number and placement of
758   EEG electrodes for measurement of neural tracking of speech 2. *bioRxiv*:800979. DOI:
759   10.1101/800979.
760 Narayanan AM, Bertrand AA. 2018. The effect of miniaturization and galvanic separation of EEG

sensor devices in an auditory attention detection task. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Institute of Electrical and Electronics Engineers Inc., 77–80. DOI: 10.1109/EMBC.2018.8512212.

Nielsen M. *Neural Networks and Deep Learning*.

Nossier SA, Rizk MRM, Moussa ND, el Shehaby S. 2019. Enhanced smart hearing aid using deep neural networks. *Alexandria Engineering Journal* 58:539–550. DOI: 10.1016/j.aej.2019.05.006.

Pan H, Pang Z, Wang Y, Wang Y, Chen L. 2020. A New Image Recognition and Classification Method Combining Transfer Learning Algorithm and MobileNet Model for Welding Defects. *IEEE Access* 8:119951–119960. DOI: 10.1109/ACCESS.2020.3005450.

Radio books for children. *Available at http://www.radioboeken.eu/kinderradioboeken.php?lang=NL* (accessed March 27, 2021).

Rashid M, Bari BS, Hasan MJ, Razman MAM, Musa RM, Nasir AFA, Majeed APPA. 2021. The classification of motor imagery response: an accuracy enhancement through the ensemble of random subspace k-NN. *PeerJ Computer Science* 7:1–31. DOI: 10.7717/peerj-cs.374.

Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, Yang GZ. 2017. Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics* 21:4–21. DOI: 10.1109/JBHI.2016.2636665.

Sanjay HS, Hiremath B V., Prithvi BS, Dinesh PA. 2020. Machine learning based assessment of auditory threshold perception in human beings. *SN Applied Sciences* 2:1–10. DOI: 10.1007/s42452-019-1929-7.

Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y. 2013. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks.

Shao S, McAleer S, Yan R, Baldi P. 2019. Highly Accurate Machine Fault Diagnosis Using Deep Transfer Learning. *IEEE Transactions on Industrial Informatics* 15:2446–2455. DOI: 10.1109/TII.2018.2864759.

Shirzhiyan Z, Shamsi E, Jafarpisheh AS, Jafari AH. 2019. Objective classification of auditory brainstem responses to consonant-vowel syllables using local discriminant bases. *Speech Communication* 114:36–48. DOI: 10.1016/j.specom.2019.09.003.

Simonyan K, Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR,.

Sriraam N. 2012. EEG based automated detection of auditory loss: A pilot study. *Expert Systems with Applications* 39:723–731. DOI: 10.1016/j.eswa.2011.07.064.

Tan L, Chen Y, Maloney TC, Caré MM, Holland SK, Lu LJ. 2013. Combined analysis of sMRI and fMRI imaging data provides accurate disease markers for hearing impairment. *NeuroImage: Clinical* 3:416–428. DOI: 10.1016/j.nicl.2013.09.008.

Tang C, Lee E. 2019. Hearing loss identification via wavelet entropy and combination of Tabu search and particle swarm optimization. In: *International Conference on Digital Signal Processing, DSP*. Institute of Electrical and Electronics Engineers Inc.,. DOI: 10.1109/ICDSP.2018.8631839.

Thorpe B, Dussard T. 2018. Classification of Speech Using MATLAB and K-Nearest Neighbour Model: Aid to the Hearing Impaired. In: *Conference Proceedings - IEEE SOUTHEASTCON*. Institute of Electrical and Electronics Engineers Inc.,. DOI: 10.1109/SECON.2018.8479223.

807   Xue P, Bai J, Wang Q, Zhang X, Feng P. 2018. Analysis and classification of the nasal finals in
808   hearing-impaired patients using tongue movement features. *Speech Communication* 104:57–
809   65. DOI: 10.1016/j.specom.2018.09.008.
810   Yan R, Gao RX, Chen X. 2014. Wavelets for fault diagnosis of rotary machines: A review with
811   applications. *Signal Processing* 96:1–15. DOI: 10.1016/j.sigpro.2013.04.015.
812   Zhang R, McAllister G, Scotney B, McClean S, Houston G. 2006. Combining wavelet analysis
813   and Bayesian networks for the classification of auditory brainstem response. *IEEE*
814   *Transactions on Information Technology in Biomedicine* 10:458–467. DOI:
815   10.1109/TITB.2005.863865.
816

Figure 1

The overall procedure of hearing deficiency diagnosis method.

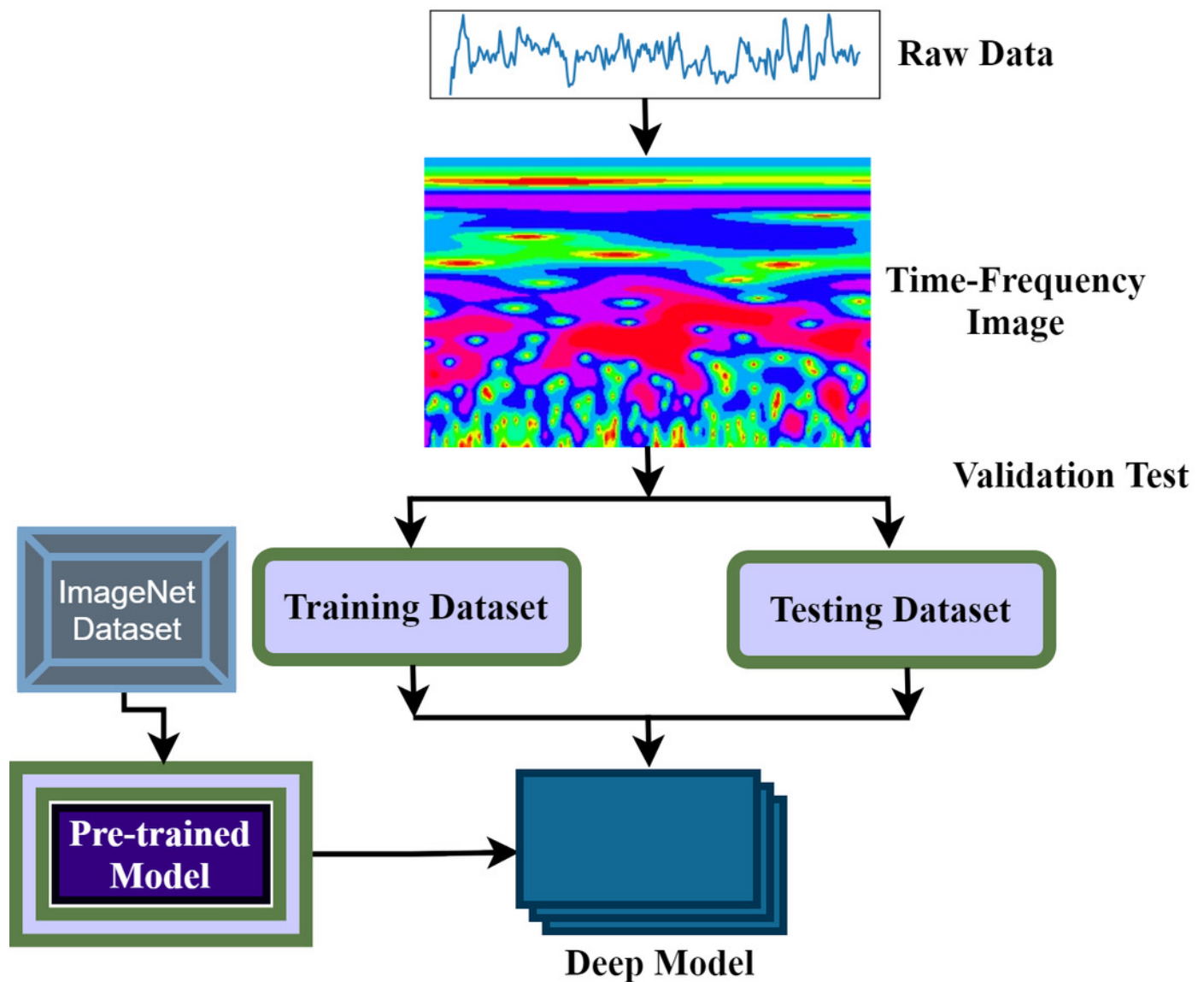



Figure 2

AEP raw data plotting in 2s decision window: (A) hear auditory stimulus with the left ear
(B) hear auditory stimulus with the right ear.

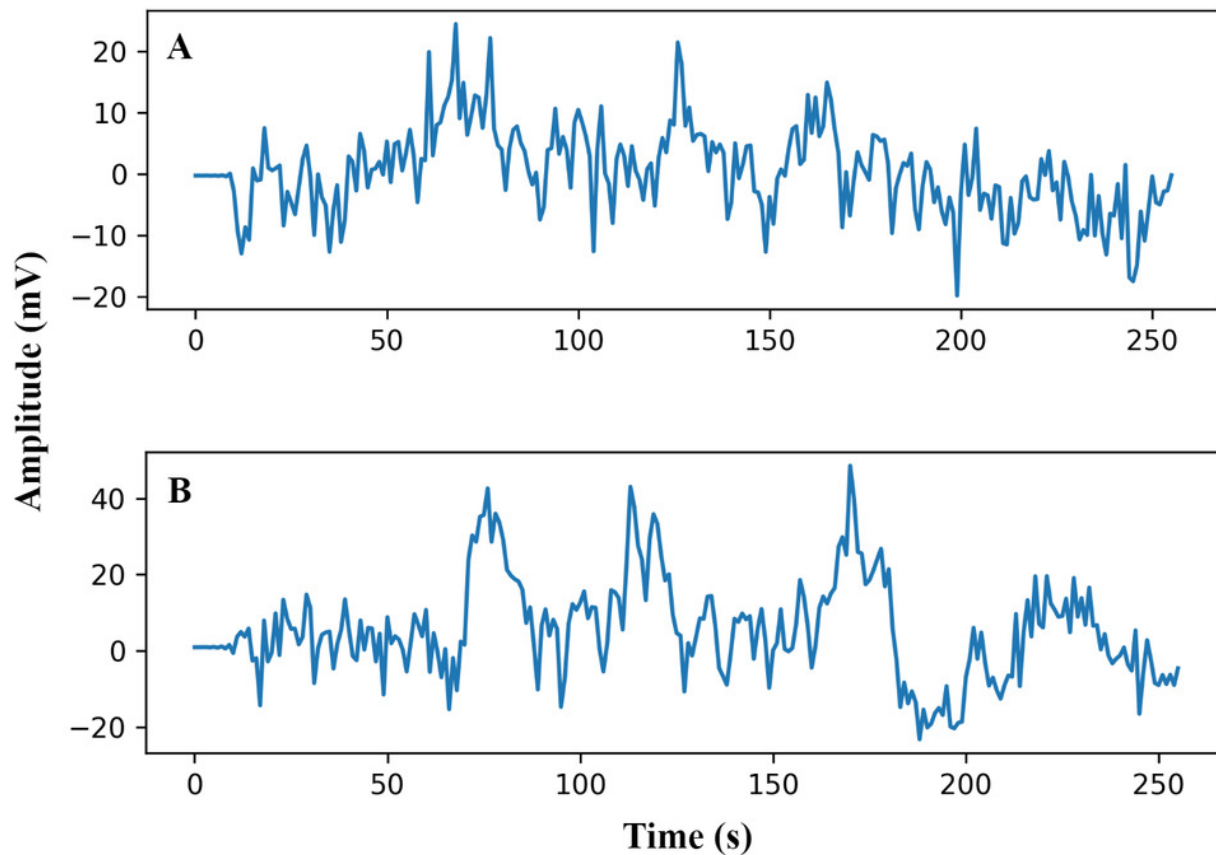



Figure 3

The Transformation process from time-domain signal to time-frequency domain image.

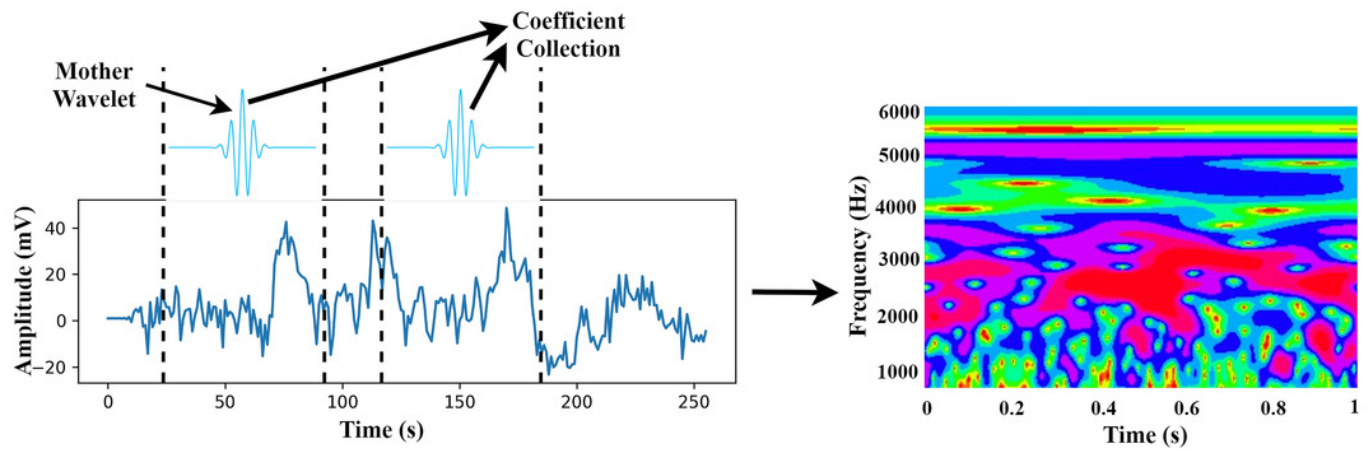



Figure 4

The time-frequency image of 64 channels data.

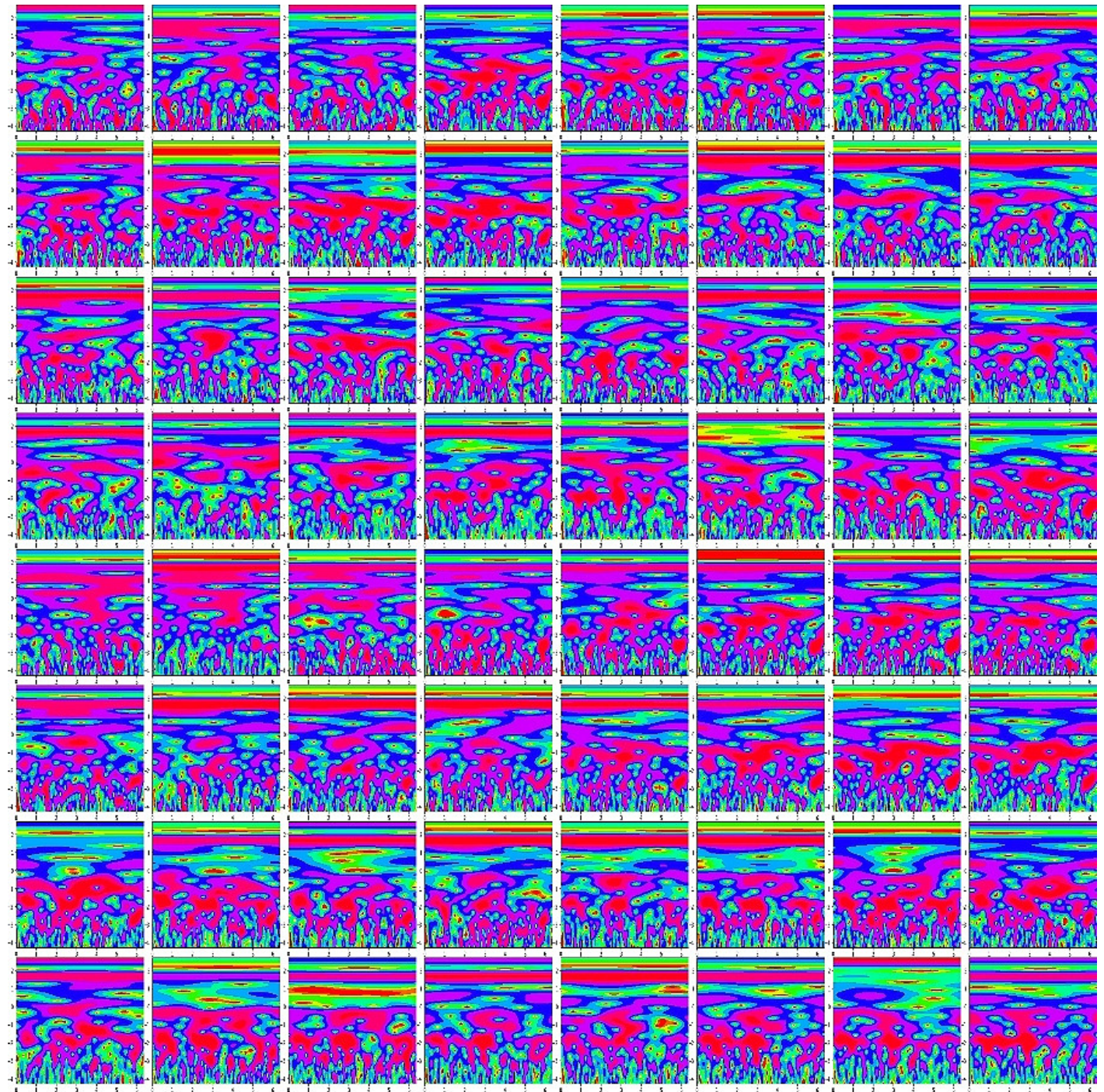



Figure 5

Typical Convolutional Neural Network Architecture.

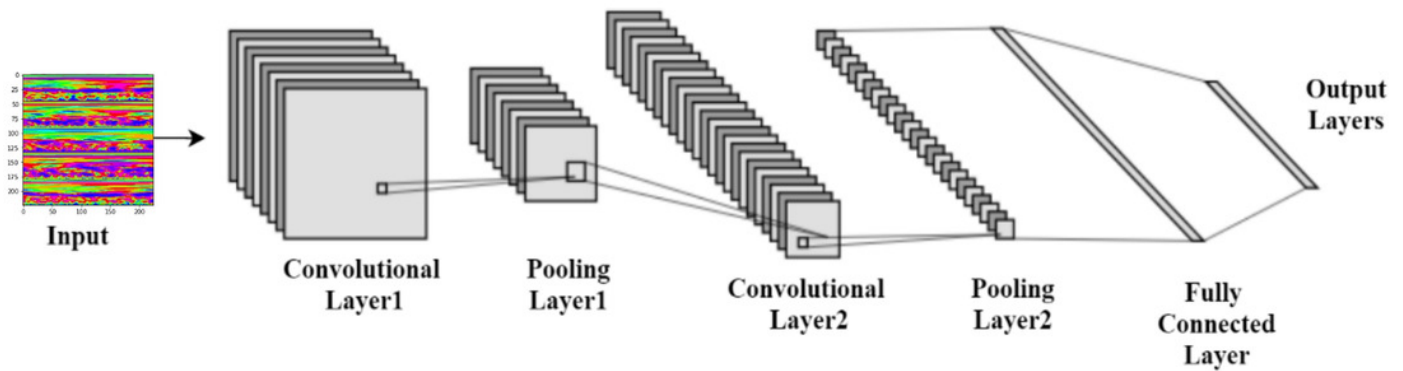



Figure 6

Transfer Learning Procedure of the Proposed Method.

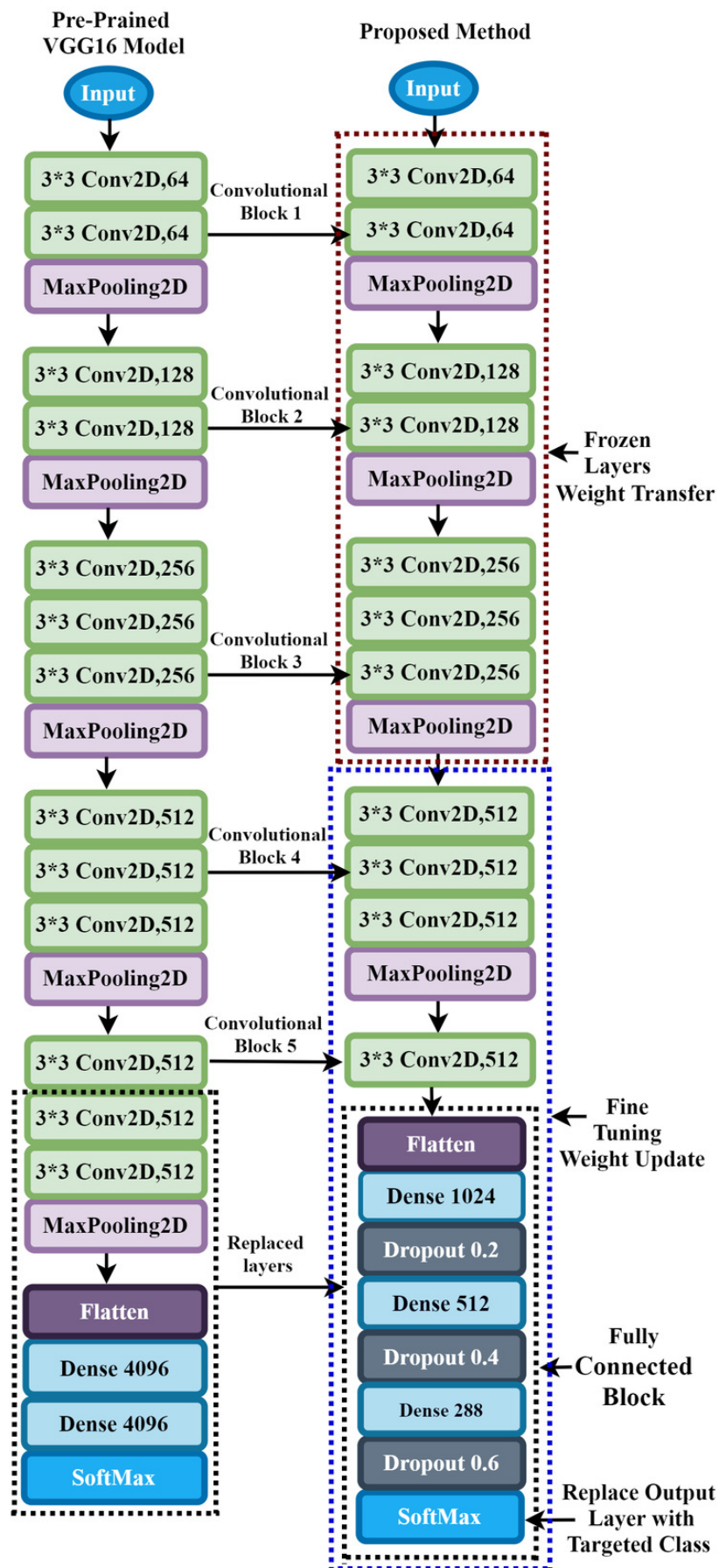



Figure 7

The overall accuracy and loss curve of the proposed TL method for 1s decision window.

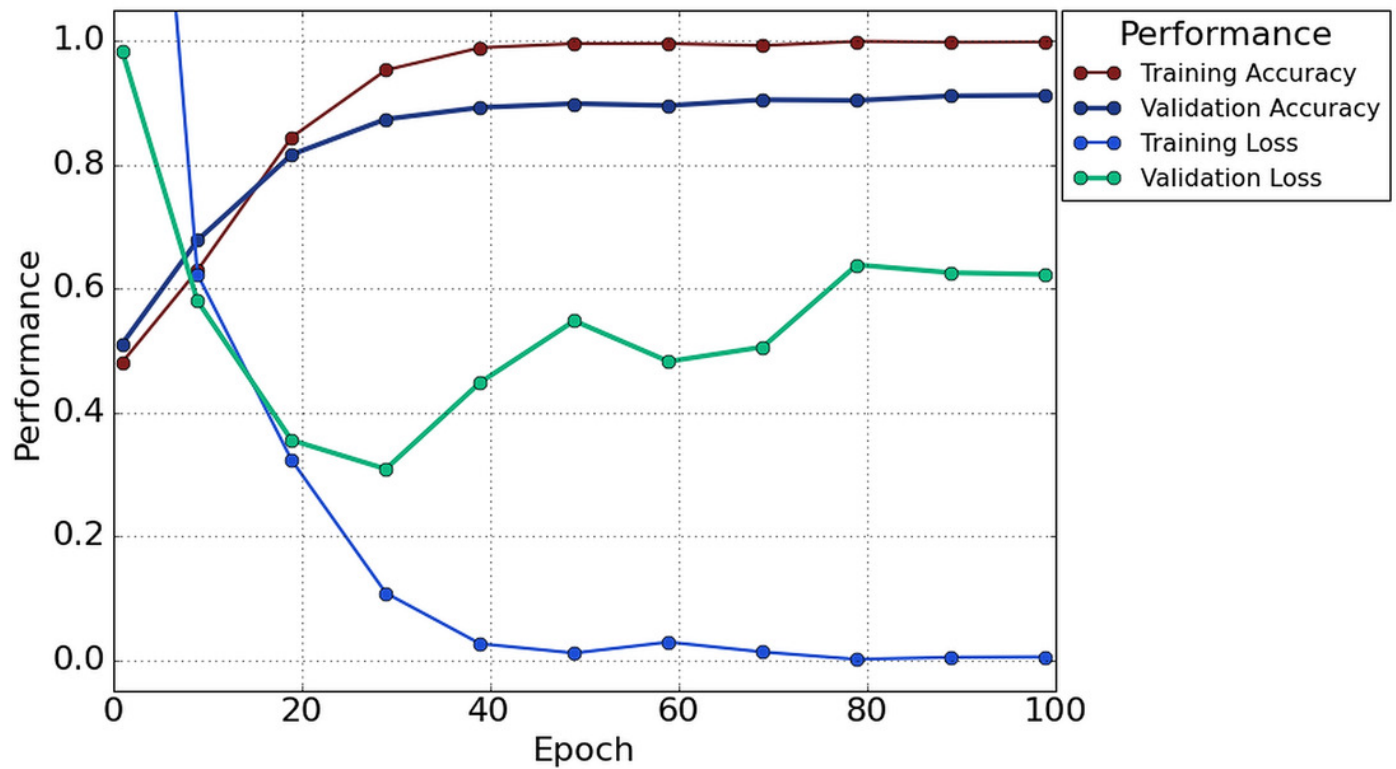



Figure 8

The overall accuracy and loss curve of the proposed TL method for 2s decision window.

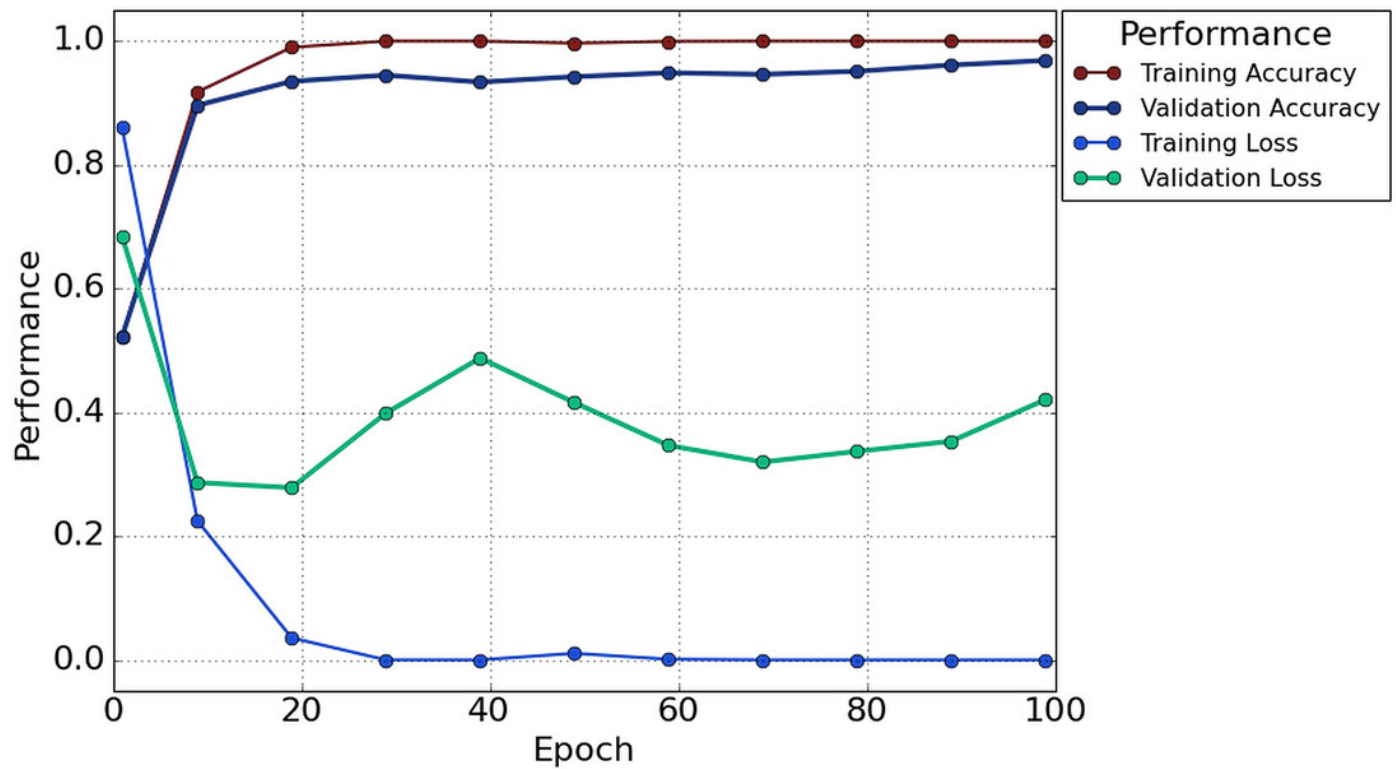



Figure 9

Confusion matrix for 1s decision windows (A)subject1 (B)subject2 (C)subject3 (D)subject4 (E)subject5 (F)subject6 (G)subject7 (H)subject8 (I)subject9 (J)subject10 (K)subject11 (L)subject12 (M)subject13 (N)subject14 (O)subject15 (P)subject16

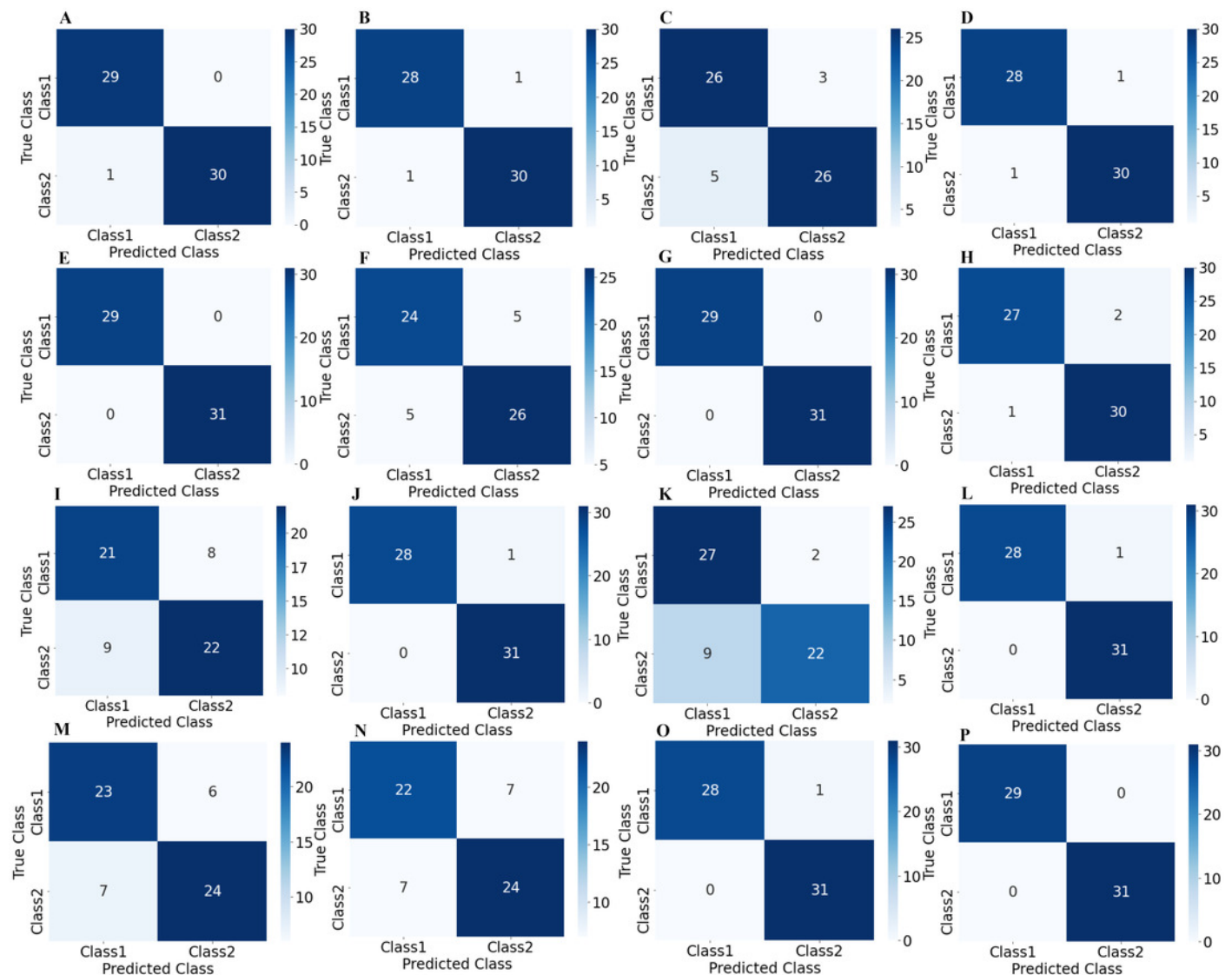



Figure 10

Confusion matrix for 2s decision windows (A)subject1 (B)subject2 (C)subject3 (D)subject4 (E)subject5 (F)subject6 (G)subject7 (H)subject8 (I)subject9 (J)subject10 (K)subject11 (L)subject12 (M)subject13 (N)subject14 (O)subject15 (P)subject16

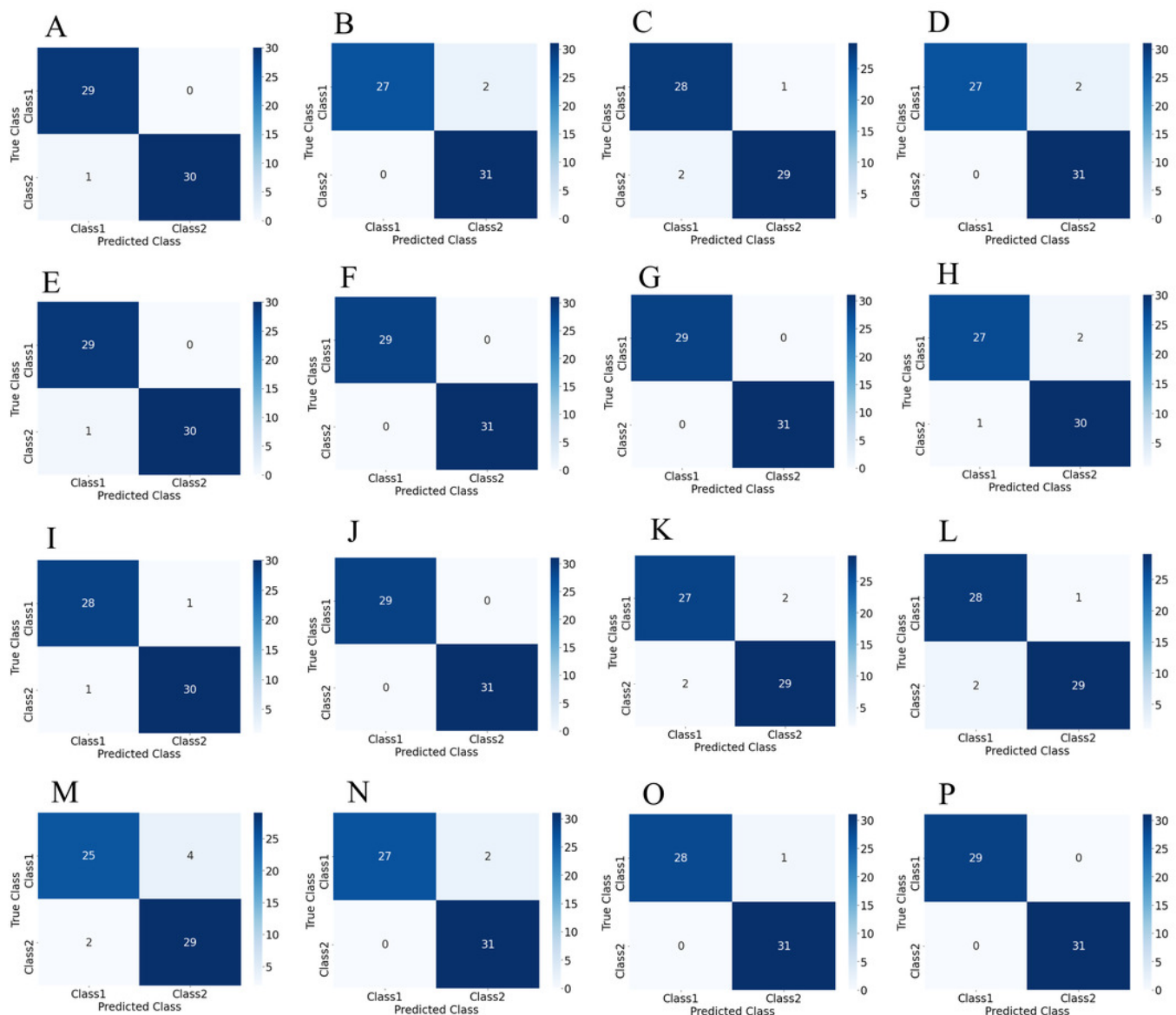



Figure 11

Hearing deficiency detection performance of the proposed TL architecture for two different window lengths.

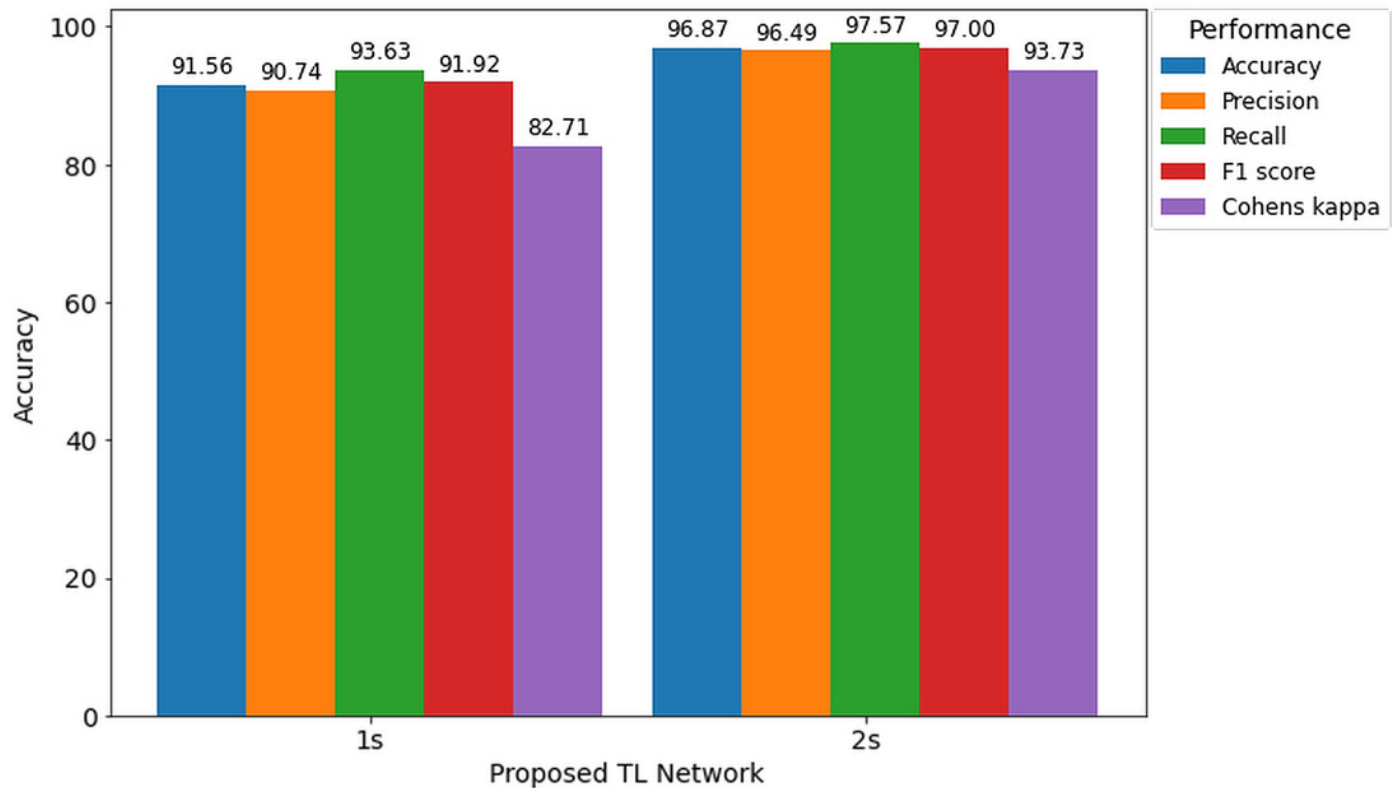
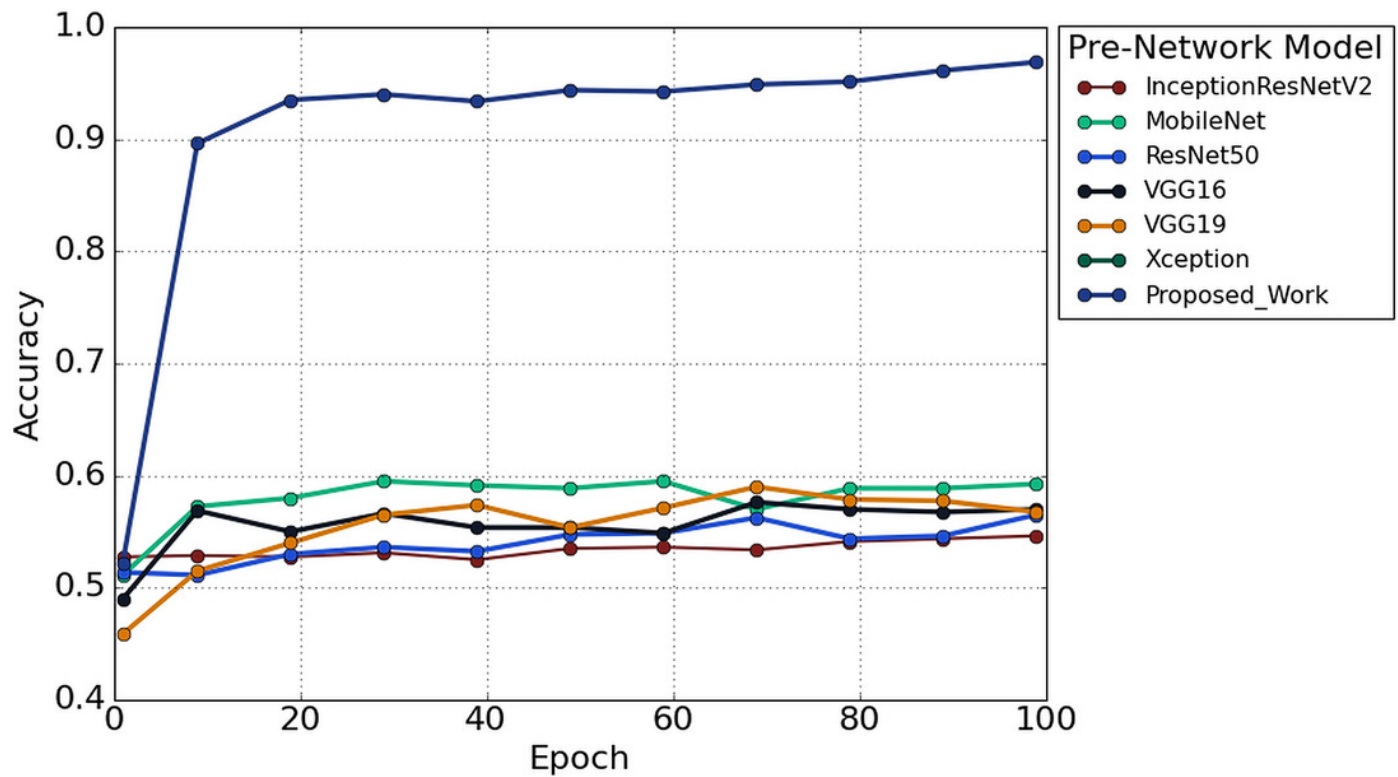



Figure 12

The performance comparison with other pre-trained architectures.

Box 1(on next page)

Training procedure of proposed TL architecture.

1   Box 1: Training procedure of proposed TL architecture.

The training steps of the proposed TL architecture:	
Step1:	Load the VGG16 base model with the pre-trained weights.
Step 2:	Freeze some layers in the base model by setting trainable = False. In the non-trainable layers, the weights will not train.
Step 3:	Create a new model by replacing some layers of VGG16 with new layers and re-train the layers with the layers where the trainable = True.
Step 4:	Train the new model with the dataset.

Parameter of proposed TL architecture.

1 Table 1: Parameter of proposed TL architecture.

Layer (type)	Output	Number of parameters
Input	224 * 224 * 3	0
Block1-Conv2D	224 * 224 * 64	1792
Block1-Conv2D	224 * 224 * 64	36928
Block1-MaxPooling2D	112 * 112 * 64	0
Block2-Conv2D	112 * 112 * 128	73856
Block2-Conv2D	112 * 112 * 128	147584
Block2-MaxPooling2D	56 * 56 * 128	0
Block3-Conv2D	56 * 56 * 256	295168
Block3-Conv2D	56 * 56 * 256	590080
Block3-Conv2D	56 * 56 * 256	590080
Block3-MaxPooling2D	28 * 28 * 256	0
Block4-Conv2D	28 * 28 * 512	1180160
Block4-Conv2D	28 * 28 * 512	2359808
Block4-Conv2D	28 * 28 * 512	2359808
Block4-MaxPooling2D	14 * 14 * 512	0
Block5-Conv2D	14 * 14 * 512	2359808
Flatten-Flatten	1 * 1 * 100352	0
fc1-Dense	1*1*1024	102761472
dropout-Dropout	1*1*1024	0
fc2-Dense	1*1*512	524800
dropout_1-Dropout	1*1*512	0
Fc3-Dense	1*1*288	147744
dropout_2-Dropout	1*1*288	0
Output-Dense	C	288*C+C

Performance of proposed model for 1s decision window.

1   Table 2. Performance of proposed model for 1s decision window.

Subject	Accuracy	Precision	Recall	F1 Score	Cohens Kappa
Subject-1	0.9833	0.9688	1.0	0.9841	0.9666
Subject-2	0.9667	1.0	0.9355	0.96667	0.9334
Subject-3	0.8667	0.8108	0.9677	0.8824	0.7312
Subject-4	0.9667	0.9393	1.0	0.9688	0.9331
Subject-5	1.0	1.0	1.0	1.0	1.0
Subject-6	0.8333	0.8387	0.8387	0.8387	0.6663
Subject-7	1.0	1.0	1.0	1.0	1.0
Subject-8	0.95	0.9667	0.9355	0.9508	0.9
Subject-9	0.7667	0.7167	0.7933	0.7367	0.4833
Subject-10	0.9833	1.0	0.9677	0.9836	0.9667
Subject-11	0.8167	0.7409	1.0	0.8578	0.6241
Subject-12	0.9833	1.0	0.9632	0.9853	0.9567
Subject-13	0.7833	0.8214	0.7419	0.7797	0.5676
Subject-14	0.76667	0.7453	0.8365	0.7892	0.5378
Subject-15	0.9833	0.9688	1.0	0.9841	0.9666
Subject-16	1.0	1.0	1.0	1.0	1.0
Average ±SD	91.56% ± 8.91%	90.74% ± 10.47%	93.63% ± 8.25%	91.92% ± 8.79%	82.71% ± 18.34%

Performance of proposed model for 2s decision window.

1 Table 3: Performance of proposed model for 2s decision window.

Subject	Accuracy	Precision	Recall	F1 Score	Cohens Kappa
Subject-1	0.9833	1.0	0.9677	0.9836	0.9666
Subject-2	0.9666	0.9393	1.0	0.9687	0.9331
Subject-3	0.95	0.9666	0.9354	0.9508	0.9
Subject-4	0.9666	0.9393	1.0	0.96875	0.9331
Subject-5	0.9833	1.0	0.9677	0.9836	0.9666
Subject-6	1.0	1.0	1.0	1.0	1.0
Subject-7	1.0	1.0	1.0	1.0	1.0
Subject-8	0.95	0.9375	0.9677	0.9523	0.8997
Subject-9	0.9666	0.9677	0.9677	0.9677	0.9332
Subject-10	1.0	1.0	1.0	1.0	1.0
Subject-11	0.9333	0.9354	0.9354	0.9354	0.8665
Subject-12	0.95	0.9666	0.9354	0.9508	0.9
Subject-13	0.9	0.8787	0.9354	0.9062	0.7993
Subject-14	0.9666	0.9393	1.0	0.9687	0.9331
Subject-15	0.9833	0.9687	1.0	0.9841	0.9665
Subject-16	1.0	1.0	1.0	1.0	1.0
Average ±SD	96.87% ± 2.78%	96.49% ± 3.5%	97.57% ± 2.76%	97% ± 2.64%	93.73% ± 5.57%

Performance comparison with six popular TL models.

1   Table 4: Performance comparison with six popular TL models.

Pre-network Model	Input Size	Trainable Parameters	Non-trainable Parameters	Total Parameters	Recognition Accuracy (%)
VGG16	224	8,194	134,260,544	134,268,738	57.375
InceptionResNetV2	224	3,074	54,336,736	54,339,810	54.000
ResNet50	224	4,098	23,587,712	23,591,810	54.875
MobileNet	224	2,002	4,253,864	4,255,866	60.250
Xception	224	4,098	20,861,480	20,865,578	57.625
VGG19	224	8,194	139,570,240	139,578,434	56.625
Proposed Model	**224**	**103,434,594**	**9,995,072**	**113,429,666**	**96.87 ± 2.78**

Performance comparison of related AEP studies.

1   Table 5: Performance comparison of related AEP studies.

Reference	Year	Data		Feature Extraction	Classification Method	Classification Accuracy (%)
		Subject	Class			
[15]	2018	180	2	WE	TS-PSO	86.17
[32]	2018	32	2	Global and nodal graph	SVM	85.71
[14]	2004	200	3	WPT	SVM	74.7
[18]	2006	8	2	DWT	Bayesian network classification	78.80
[33]	2013	39	2	SIFT	SVM	87
[19]	2019	Observation: 671	2	FFT	SVM	78.7
[34]	2019	Observation:671	2	Raw AEP	CNN	94.1
[35]	2016	Observation: 280 Subjects: 151	2	latency, FFT and DWT	A feed-forward multilayer perceptron	90.74
Proposed Work	-	**Observations: 3200**	2	**CWT**	**Improved-VGG16**	**96.87**