Classification of Known and Unknown Study Items in a Memory Task using Single-Trial Event-Related Potentials and Convolutional Neural Networks

Jorge Armando Delgado-Munoz * , Reiko Matsunaka , Kazuo Hiraki

Posted Date: 17 July 2024

doi: 10.20944/preprints202407.1239.v1

Keywords: Long term memory; Familiarity; Electroencephalography; Event-Related Potentials; Convolutional Neural Networks

 doi:10.20944/preprints202407.1239.v1

Article

Classification of Known and Unknown Study Items in a Memory Task using Single-Trial Event-Related Potentials and Convolutional Neural Networks

Jorge Delgado-Munoz *, Reiko Matsunaka and Kazuo Hiraki [1]

Graduate School of Arts and Sciences, The University of Tokyo, Meguro-Ku Tokyo 153-8902, Japan; hiraki-lab@ardbeg.c.u-tokyo.ac.jp

* Correspondence: joardemu@gmail.com

Abstract: This study examines the feasibility of using event-related potentials (ERPs) obtained from electroencephalographic (EEG) recordings as biomarkers for long-term memory item classification. Previous studies have identified old/new effects in memory paradigms associated with explicit long-term memory and familiarity. Recent advancements in convolutional neural networks (CNNs) have enabled the classification of ERP trials under different conditions and identification of features related to neural processes at the single-trial level. We employed this approach to compare three CNN models with distinct architectures using experimental data. Participants (N = 25) performed an association memory task while recording ERPs that were used for training and validation of the CNN models. The EEGNET-based model achieved the most reliable performance in terms of precision, recall, and specificity compared with the shallow and deep convolutional approaches. The classification accuracy of this model reached 70% for known items and 62% for unknown items. Good overall accuracy requires a trade-off between recall and specificity and depends on the architecture of the model and the dataset size. These results suggest the possibility of integrating ERP and CNN into online learning tools and identifying the underlying processes related to long-term memorization.

Keywords: long term memory; familiarity; electroencephalography; event-related potentials; convolutional neural networks

1. Introduction

Long-term memory can be classified into two types based on the type of information that must be retrieved: implicit and explicit. Implicit memory is related to procedures and the execution of specific tasks; it is not retrieved consciously and is mostly associated with skills or daily tasks that do not require relearning them to be performed. Explicit memory, in contrast, refers to the storage of factual and objective information through textbook learning or experiential memories, which are commonly acquired through rehearsal and must be retrieved consciously according to when such information is needed [1,2]. Explicit memory can be classified into episodic memory, which is related to specific events, and semantic memory, which is related to facts, concepts, and general knowledge.

Yonelinas [3] distinguished between two fundamental processes of long-term memory retrieval: recollection and familiarity. Recollection refers to the retrieval of details and contextual information regarding past events that surround the acquisition of new knowledge and memories. Alternatively, familiarity is based on the qualitative signal of an item, and commonly refers to the capacity to identify an item by its name without any background or contextual information [4]. Distinguishing familiarity and recollection-based recognition remains a complex endeavor because these processes are associated with distinct brain regions and cognitive mechanisms [4].

Since the late 1990s, event-related neuroimaging has been employed to investigate the impact of different encoding strategies on subsequent memory retrieval, establishing an early foundation for the connection between familiarity and neural processes [5]. Several studies conducted since the beginning of the new millennium have used the event-related potential (ERP) technique as a tool for

 doi:10.20944/preprints202407.1239.v1

elucidating memory encoding and retrieval processes, emphasizing its high temporal resolution for studying the timing of the cognitive processes involved in memory [6]. Around the mid-2000s, advances in research yielded significant insights into the electrophysiological aspects of familiarity, with extensive studies exploring the FN400 component, an ERP response consistently linked to familiarity, and the old/new effect [7–9]. Recent evidence suggests that FN400 is sensitive to changes in contextual familiarity, indicating its role in the comparative evaluation of familiarity within a specific context [10]. The N400 component, typically associated with semantic processing, has also been implicated in familiarity assessments [9,10]. N400 appears to reflect an evaluation of absolute familiarity, independent of contextual information, introducing an additional layer of complexity to our understanding of the electrophysiological responses associated with familiarity. The last decade has witnessed a surge in the volume of studies dedicated to understanding familiarity and its neural substrates. This growing body of literature provides profound insights into the neural mechanisms underlying recognition memory, particularly focusing on the distinction between familiarity-based recognition and recollection-based source memory retrieval [11]. However, individual variations in reliance on familiarity versus recollection processes during recognition memory tasks have been highlighted, underscoring the non-uniformity of these processes across individuals [12].

One typical situation in which explicit semantic memory is applied is the memorization of a set of items related to a specific subject as part of an evaluation and assessment in different learning and education contexts. Such a set might contain a few items already known by the student that do not require restudy and others that the student may encounter for the first time or might not be recalled with sufficient confidence. This opens up the possibility of using the electrophysiological response as a biomarker to predict whether an item has been previously encountered or learned. Studies on the use of electrophysical recordings to quantitatively predict the degree of familiarity with a determined item during memorization tasks have been conducted. Fukuda and Woodman [13] conducted a study that specifically focused on predicting the recognition memory of an individual using electroencephalographic (EEG) signals. They collected EEG signals from participants engaged in recognition memory tasks. Their study focused on two key neural signals: the P3 component of the ERP and alpha power modulation. The latter, which refers to changes in the power of neural oscillations in the alpha frequency band (8– 12 Hz), is assumed to be indicative of the inhibitory processes involved in gating sensory information and has been associated with memory performance. Furthermore, their study explored potential real-time interventions that could enhance memory performance based on predictions from EEG signals. Similarly, Khurana et al. [14] conducted a study aimed at investigating the combination of different EEG features and frequency bands to accurately predict word familiarity. The study involved recording EEG signals from participants as they were presented with words of varying familiarity levels. Both time- and frequency-domain features were extracted and analyzed. This study determined that a combination of specific EEG features and frequency bands resulted in an accurate prediction of word familiarity. This suggests a strong correlation between these EEG features and frequency bands and the familiarity ratings of words, further emphasizing the importance and potential of EEG features in predicting cognitive states.

Typically, ERP analysis is performed by epoching time-locked segments of the recorded EEG signal around the stimulus onset, referred to as trials, grouped into different categories depending on the condition of the task or the behavioral response of the participant, and thereafter averaging the signals from each group such that the signal-to-noise ratio of the recording is increased, and the components related to the neural process associated with the task are highlighted while canceling the irrelevant ones during the averaging operation [15–17]. Although this method usually achieves satisfactory results, the use of ERP in memory tasks has certain peculiarities that make it inadequate [18]. Memory paradigms have an implicit imbalance in the number of items that will be remembered or recognized, making the class with fewer trials not to contain the necessary number of items required to identify components associated with the process. Moreover, the ability to encode new information is not perfect and varies from one individual to another [17]. This issue requires the

 doi:10.20944/preprints202407.1239.v1

analysis of ERP signals at the single-trial level, allowing the elucidation of existing individual variability [18].

To address these limitations, researchers have proposed various more sophisticated methods for single-trial analysis of EEG and ERP for classification at the single-trial level. Deep learning (DL) methodologies, including convolutional neural networks (CNNs) and recurrent neural networks, are of interest. In particular, CNNs, which have proven to be effective in analyzing image data, have achieved promising results in processing physiological signal datasets, such as epoched EEG recordings. Researchers have adapted these network architectures to process EEG data that involve time-series information representing neural activity over time [19–22]. CNNs have reportedly achieved a high classification accuracy in distinguishing between different types of stimuli and diagnoses. CNNs offer promising solutions for automated feature extraction and direct classification using raw EEG data [23]. This automation can overcome the challenges associated with feature engineering and selection that are inherent in traditional methods. Finally, the flexibility of neural networks in modeling nonlinear relationships and complex interactions within ERP data enables them to capture intricate patterns and dependencies. Their generalization capability enhances the applicability and reliability of single-trial analysis in diverse research settings, including rapid and on-the-fly analysis in experimental paradigms, such as brain– computer interfaces (BCIs) and neurofeedback applications [24]. DL methods have been successfully applied to classification problems involving time-locked stimuli in EEG recordings and ERP, such as the recognition of rhythm pattern perception [19], seizure detection [23], diagnosis of schizophrenia [20], neuromarketing [25], and attention levels during driving [26]. Using the proposed approach, we investigate the feasibility of integrating devices capable of recording and analyzing electrophysiological responses into digital and online learning tools to adjust the volume of learning items. The ERP response elicited by the presentation of each study item is used as the input of a CNN model to predict whether such an item has already been learned by the student and excluded from the learning list, or whether it is a newly encountered item that must be learned through repetition.

2. Materials and Methods

2.1. Participants

A group of 25 participants (N = 25) was recruited for the experiment. All participants were students at the University of Tokyo with an age range of 18–30 years (mean age = 20.96, SD = 3.06); 68% were male and 32% were female. Prior to the task, all participants were asked whether they were interested in geography and to decide whether to perform the task in Japanese or English; none of the participants reported any history of psychiatric disease. Data from four participants were excluded from the analysis owing to one of them reporting color blindness after the execution of the task, two of them owing to excessive signal artifacts and one of them owing to corrupted behavioral data. Moreover, the data from three of the aforementioned participants correspond to individuals who did not perform the task in their native languages. The final analysis was performed on data from the remaining 21 participants (N = 21) who completed the task in Japanese (mean age = 20.04, SD = 2.13; 71.43% male, 28.57% female). The experiment was conducted in accordance with the Declaration of Helsinki and the ethics regulations of The University of Tokyo. Informed consent was obtained from all participants involved in the study, and they received monetary compensation for their cooperation.

2.1. Experimental Design

The experiment was designed using JSPsych, a JavaScript library specially designed to conduct psychology experiments capable of running on a web browser [27,28]. The experiment was conducted by presenting the task on a color 23-inch LCD computer screen in a soundproof room with controlled temperature.

The experimental task comprised a full study session divided into three sections: pretest, encoding, and test, similar to the study conducted by Fukuda and Woodman [13]. The comprehensive

 doi:10.20944/preprints202407.1239.v1

structure of the task designed in this study is shown in Figure 1. The experiment began by providing a practice section to familiarize participants with the elements on the screen. Prior to the start of each section, a set of instructions was displayed on the screen and the participants were required to click on the START button to initiate the experiment. Each item presented on the screen comprised a display of a fixation cross for a variable time (800, 900, 1000, 1200, or 1300 ms) followed by the presentation of the visual stimuli with a prompt. The item set comprised the flags of United Nations member countries. For each participant, a subset of 60 items was randomly selected from the entire set of 193. During the pretest section, the prompt asked the participants whether they knew the name of the country to which the displayed flag belongs and required them to click the button corresponding to their desired answer (YES or NO). After the presentation of all the items, the screen displayed a message inviting the participants to take a short break before continuing with the next section. The encoding section involved practicing the entire list in a spaced repetitive manner. The set was divided into four subsets of 15 elements each, showing a flag for 1250 ms, followed by the corresponding country name. When the name of the country was displayed, the participant was invited to click the NEXT button to proceed to the next item. Once all 60 items had been studied, the participants were required to take a short break before studying the items three more times. Each time, the order of the item presentation was randomly changed. Finally, a test was performed to evaluate the number of items remembered after the encoding section. In this section, the flag of the country was displayed first, followed by a series of prompts on which the participants were asked to click a button whether they could remember the number of the country (NO or YES) and whether they felt confident about their answer (CONFIDENT, JUST GUESSING). Subsequently, the correct name of the country was displayed, and the final prompt asked the participants whether it was the name of the country they remembered. This sequence of prompts is conditioned by the first response. The participants were instructed to carefully read all the prompts and indications displayed on the screen during the task and blink only when a fixation cross appears on the screen.

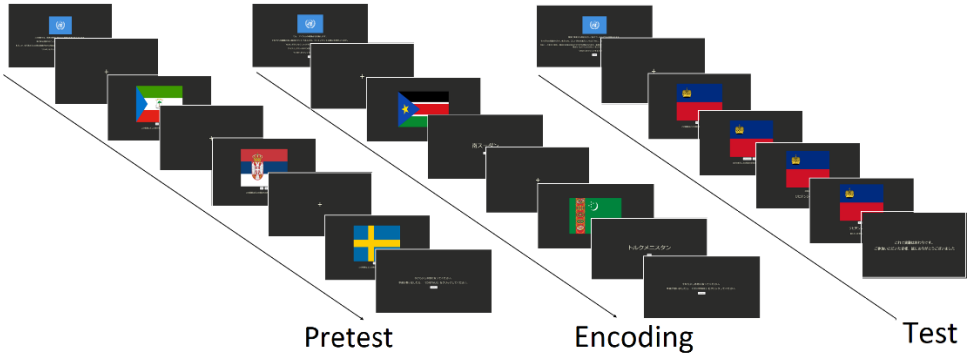


Figure 1. Structure of the experimental task. Participants were instructed to assess their previous knowledge during the pretest section, study the items during the encoding section, and evaluate the number of items they could remember in the test section. Only the pretest section is relevant for this article.

2.2. EEG Signal Recording

EEG signals were collected using a Geodesic EEG system (Magstim EGI Inc., Eugene, OR, USA) comprising a Geodesic Sensor Net with 128 Ag/AgCl wet electrodes, a Net Amps 400 medical grade biosignal amplifier, and the Net Station software suite. The signals were recorded at a sampling rate of 250 Hz and referenced online to Cz. Before starting the signal recording, the impedance of each electrode was maintained under 10 KΩ.

Events corresponding to item presentation on the screen were captured using a Cedrus StimTracker system (Cedrus Corporation, San Pedro, CA, USA). The stimuli were captured using StimTracker's optical sensor attached to the LCD screen and synchronized with the EEG recording at the NET station via a parallel port connection between the StimTracker device and the amplifier. A schematic of the experimental setup is shown in Figure 2.

 doi:10.20944/preprints202407.1239.v1

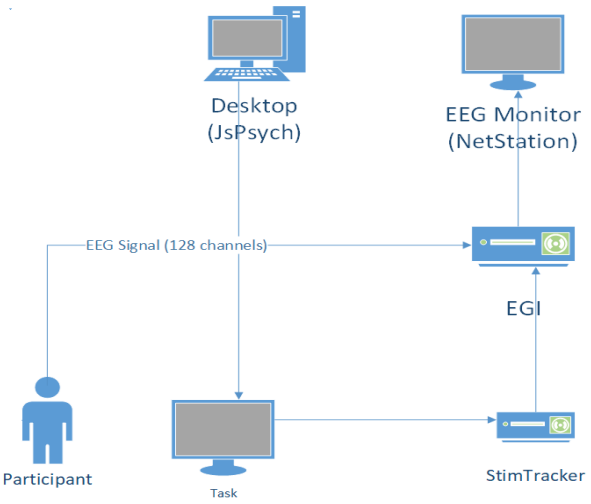


Figure 2. Experimental setup schematic. This diagram describes the apparatus and their use during the experimental task.

2.3. EEG Signal Preprocessing

EEG signal preprocessing was performed using MATLAB R2023a (Mathworks Inc., Natick, MA, USA). The recording archives were first imported using EEGLAB [29], an open-source toolbox for MATLAB used for the processing and analysis of neural data. Following the indications of Calbi et al. [30], the channels corresponding to the outer belt were removed prior to preprocessing to avoid the presence of muscle and movement artifacts, thereby reducing the initial number of electrodes from 128 to 110. Subsequently, the entire EEG recording archives for each participant were segmented into three parts corresponding to different sections of the task.

Each recording segment was processed as follows: Signals were filtered using a notch filter at 50 Hz to remove power line noise, and a bandpass filter between 0.1 and 30 Hz to eliminate DC offset and EMG artifacts. The EEG channels were re-referenced offline to the average of the left and right mastoids, allowing reconstruction of the Cz channel. Subsequently, the dataset was downsampled to 128 Hz to comply with the input requirements for generating the CNN models. The remaining channels were clustered according to the 10–20 system to reduce the number of channels from 110 to 22 [30,31]. The criteria for performing this clustering were to select the channel labeled according to the 10–20 system as a reference and to average it with adjacent channels. Channels that show excessive artifacts were excluded from their corresponding clusters, and only the cleanest channels were used. Figure 3 shows the channels that were initially removed and the clustering of different electrode groups to reduce the number of channels.

 doi:10.20944/preprints202407.1239.v1

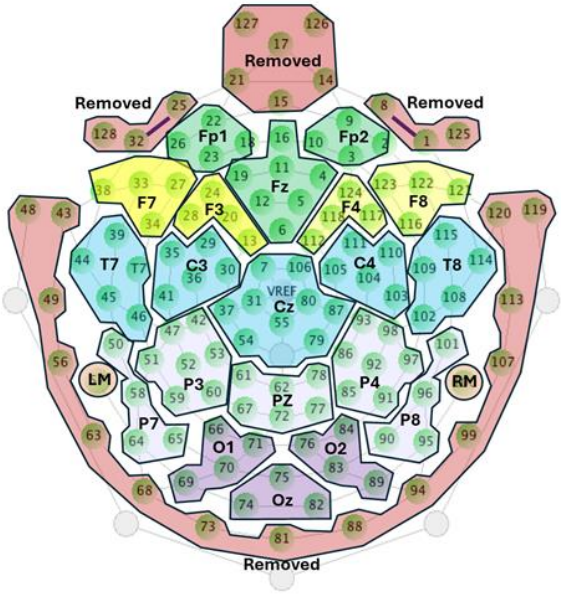


Figure 3. Clustering of electrodes for channel number reduction. Channels marked in red were removed prior to signal preprocessing.

2.4. ERP Processing

The ERPLAB plugin [33] of EEGLAB was used to identify the ERP responses associated with each stimulus. Once the data were preprocessed, event information containing the behavioral responses of the participants and class information were imported into the datasets corresponding to each participant. The data were epoched selecting a time window between -200 and 1000 ms from the stimulus presentation and baseline corrected at 200 ms prior to the stimulus presentation. Using a standardized measurement error tool [34], channels Fp1 and Fp2 were rejected because of the presence of excessive noise in the majority of the participants' data, reducing the number of channels from 22 to 20. Next, an artifact correction operation was performed. First, eye–eye blink artifacts were removed using ICA and the IClabel tool from EEGLAB. Subsequently, ERP segments that contained peak-to-peak artifacts and step-like artifacts were rejected using the artifact removal tool from ERPLAB. All the trials were exported to each individual CSV files, starting from stimulus onset to the duration of the previously defined trial (1000 ms / 128 datapoints). Each CSV file comprises a two-dimensional array, where the number of rows corresponds to that of EEG channels, and that of columns corresponds to that of samples in the trial [35].

2.5. CNN Architectures Description

2.5.1. EEGNet

EEGNet is a CNN architecture specifically designed for EEG signal processing in BCIs [36]. This architecture comprises a temporal convolution layer, which learns frequency filters; a depthwise convolution layer, which learns specific spatial filters connected to each feature map individually; a separable convolution layer, which combines depthwise convolution and learns a temporal summary for each feature map individually; and a pointwise convolution layer, which learns to mix the feature maps together. Figure 4 shows the structure of the NN architecture. EEGNet-based models can be customized by modifying parameters C, T, F1, and F2, where C denotes the number of channels; T denotes the number of timepoints or samples in each trial; F1 denotes the number of temporal filters; D denotes the number of spatial filters; F2 denotes the number of pointwise filters, defined as F1 × D; and N denotes the number of classes. In addition, the filter size for the first layer is defined as half

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 July 2024 doi:10.20944/preprints202407.1239.v1

the number of samples. The notation EEGNET-8,2 refers to the default parameters of the architecture, with F1 = 8 and D = 2.

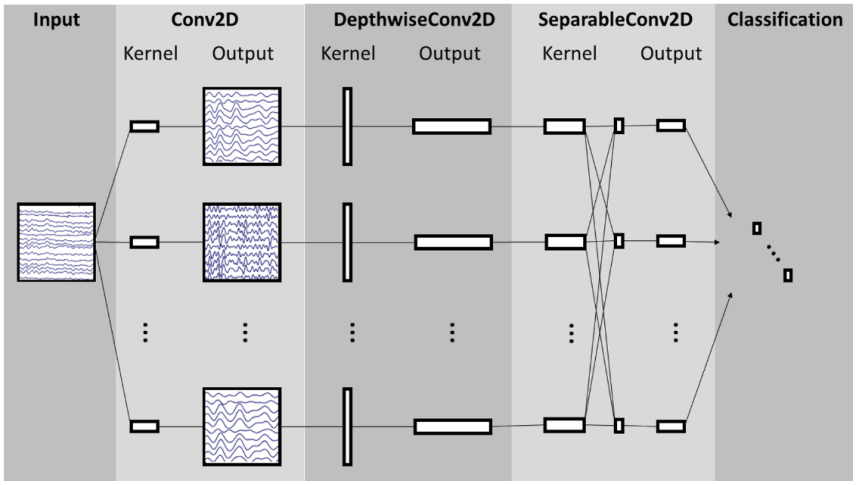
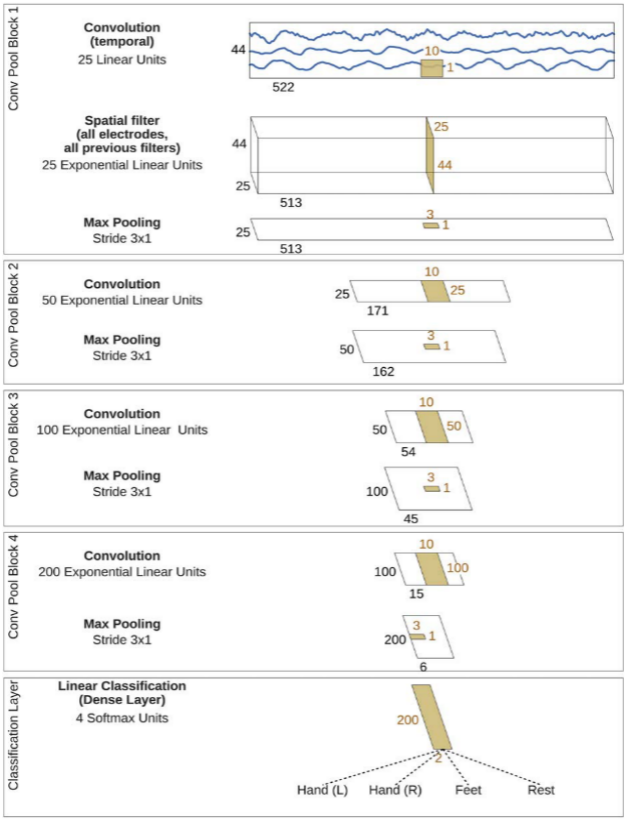


Figure 4. EEGNet architecture [36].

2.5.2. Deep Convolutional Neural Network (DeepConvNet)

The deep convolutional neural network (DeepConvNet) used in this study was conceived as a model capable of extracting a wide range of features without being limited to specific types [35]. This generic architecture aims to achieve competitive accuracies with minimal expert knowledge and demonstrates the potential of standard CNNs for brain-signal decoding tasks. The architecture comprised four convolution-max-pooling blocks, with the first block specially designed to handle the EEG input. It used two layers in the first convolutional block to handle several input channels, and the second layer performed spatial filtering across electrode pairs. Exponential linear units were used as activation functions. Design choices were evaluated against alternative options, such as rectified linear units. The basic architecture of the multiclass classification problem is illustrated in Figure 5.



 doi:10.20944/preprints202407.1239.v1

Figure 5. DeepConvNet architecture for multiclassification problem [35].

2.5.3. Shallow Convolutional Neural Network (ShallowConvNet)

The (ShallowConvNet) architecture was inspired by the FBCSP pipeline and optimized to decode band power features [35]. This architecture comprises temporal convolution and spatial filtering akin to FBCSP, with larger kernel sizes to allow for a broader range of transformations. The following steps are squaring nonlinearity, mean pooling, and logarithmic activation functions, mirroring the trial log-variance computation in FBCSP. Unlike FBCSP, ShallowConvNet integrates all the computational steps into a single network, enabling joint optimization. In addition, it incorporates multiple pooling regions within one trial, facilitating the learning of the temporal structures of band power changes, which have been shown to improve classification accuracy. The ShallowConvNet architecture is illustrated in Figure 6.

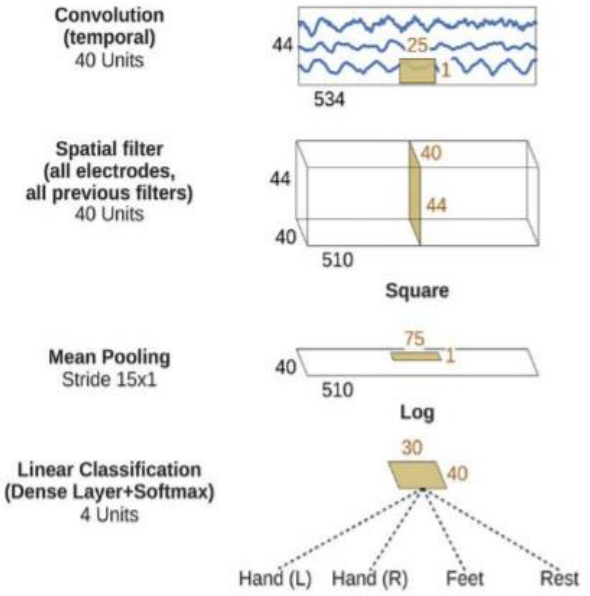


Figure 6. ShallowConvNet architecture for multi-classification problem [35].

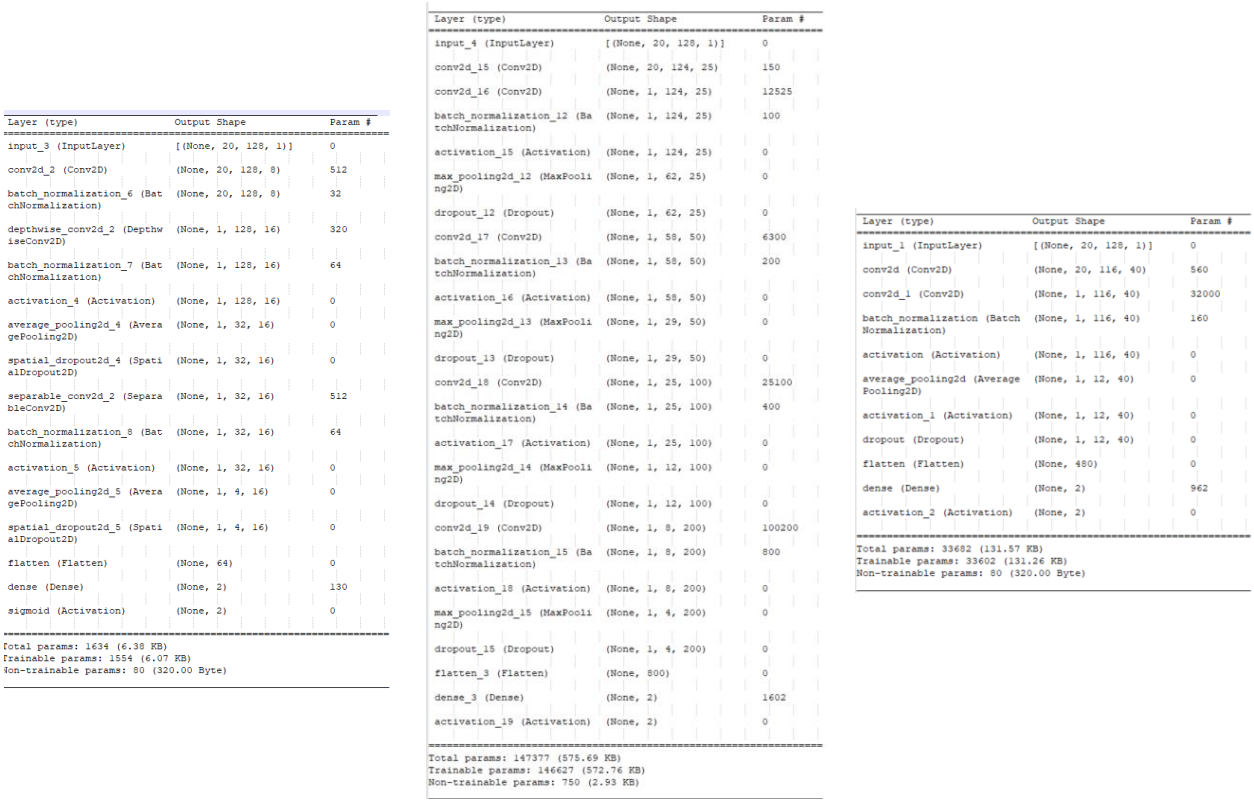
2.6. Pretraining of the Models

The data used to train the CNN models corresponded to the trials performed in the pretest section, for a total of 1260 trials. The artifact rejection operation resulted in 1153 usable and 107 rejected trials. The operating data were separated into 80% of the trials for use as training data, and the remaining 20% were used as testing data. The architectures for the models were implemented using an online library written in Python using the TensorFlow and Keras frameworks [35,37]. The library was slightly modified to use a sigmoid activation function instead of softmax in the last layer, and imported into the Google Colaboratory (Colab) environment for training and validation. This environment runs on an Intel® Xeon® CPU at 2.30 GHz, an NVIDIA Tesla T4 GPU accelerator, and 12.72 GB of RAM memory. A summary of the three architectures is shown in Figure 7.

After the data were imported into the development environment, each channel was normalized such that the signal amplitude ranged from 0 to 1. Because the input data proportion of trials labeled as "Unknown" with respect to those labeled as "Known" is approximately 2.7:1, both training and testing sets were augmented using the RandomOverSampler function from the imblearn library, using maximization of the minoritarian class in the set as the sampling strategy. This resulted in 1342 trials for the training set and 350 trials for the testing set. Additionally, the same number of trials as in the testing set were randomly extracted from the training set as validation data, leaving 60% of the total data for training the model, 20% for validation, and 20% for testing. The models were trained using an ADAM optimizer with a learning rate of 0.0001 and a binary cross-entropy loss function.

 doi:10.20944/preprints202407.1239.v1

Finally, a checkpoint callback was implemented to store the weights that provided the lowest result for the loss function at the end of the training. Model pretraining was performed using 300 epochs, a batch size of 64 for EEGNET, and a batch size of 16 for DeepConvNet and ShallowConvNet. Figure 6 shows the summary of the model architectures generated during this step. Finally, the models with the best overall performance based on the overall accuracy and loss function across the training and true positive rate for each class assessed by their respective confusion matrices were selected, and their respective weights were stored externally for further cross-validation.



Layer (type)	Output Shape	Param #
input_3 (InputLayer)	[(None, 20, 128, 1)]	0
conv2d_2 (Conv2D)	(None, 20, 128, 8)	512
batch_normalization_6 (BatchNormalization)	(None, 20, 128, 8)	32
depthwise_conv2d_2 (DepthwiseConv2D)	(None, 1, 128, 16)	320
batch_normalization_7 (BatchNormalization)	(None, 1, 128, 16)	64
activation_4 (Activation)	(None, 1, 128, 16)	0
average_pooling2d_4 (AveragePooling2D)	(None, 1, 32, 16)	0
spatial_dropout2d_4 (SpatialDropout2D)	(None, 1, 32, 16)	0
separable_conv2d_2 (SeparableConv2D)	(None, 1, 32, 16)	512
batch_normalization_8 (BatchNormalization)	(None, 1, 32, 16)	64
activation_5 (Activation)	(None, 1, 32, 16)	0
average_pooling2d_5 (AveragePooling2D)	(None, 1, 4, 16)	0
spatial_dropout2d_5 (SpatialDropout2D)	(None, 1, 4, 16)	0
flatten (Flatten)	(None, 64)	0
dense (Dense)	(None, 2)	130
sigmoid (Activation)	(None, 2)	0

Total params: 1634 (6.38 KB)
Trainable params: 1554 (6.07 KB)
Non-trainable params: 80 (320.00 Byte)

(a)

Layer (type)	Output Shape	Param #
input_4 (InputLayer)	[(None, 20, 128, 1)]	0
conv2d_15 (Conv2D)	(None, 20, 124, 25)	150
conv2d_16 (Conv2D)	(None, 1, 124, 25)	12525
batch_normalization_12 (BatchNormalization)	(None, 1, 124, 25)	100
activation_15 (Activation)	(None, 1, 124, 25)	0
max_pooling2d_12 (MaxPooling2D)	(None, 1, 62, 25)	0
dropout_12 (Dropout)	(None, 1, 62, 25)	0
conv2d_17 (Conv2D)	(None, 1, 58, 50)	6300
batch_normalization_13 (BatchNormalization)	(None, 1, 58, 50)	200
activation_16 (Activation)	(None, 1, 58, 50)	0
max_pooling2d_13 (MaxPooling2D)	(None, 1, 29, 50)	0
dropout_13 (Dropout)	(None, 1, 29, 50)	0
conv2d_18 (Conv2D)	(None, 1, 25, 100)	25100
batch_normalization_14 (BatchNormalization)	(None, 1, 25, 100)	400
activation_17 (Activation)	(None, 1, 25, 100)	0
max_pooling2d_14 (MaxPooling2D)	(None, 1, 12, 100)	0
dropout_14 (Dropout)	(None, 1, 12, 100)	0
conv2d_19 (Conv2D)	(None, 1, 8, 200)	100200
batch_normalization_15 (BatchNormalization)	(None, 1, 8, 200)	800
activation_18 (Activation)	(None, 1, 8, 200)	0
max_pooling2d_15 (MaxPooling2D)	(None, 1, 4, 200)	0
dropout_15 (Dropout)	(None, 1, 4, 200)	0
flatten_3 (Flatten)	(None, 800)	0
dense_3 (Dense)	(None, 2)	1602
activation_19 (Activation)	(None, 2)	0

Total params: 147377 (575.69 KB)
Trainable params: 146627 (572.76 KB)
Non-trainable params: 750 (2.93 KB)

(b)

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 20, 128, 1)]	0
conv2d (Conv2D)	(None, 20, 116, 40)	560
conv2d_1 (Conv2D)	(None, 1, 116, 40)	32000
batch_normalization (BatchNormalization)	(None, 1, 116, 40)	160
activation (Activation)	(None, 1, 116, 40)	0
average_pooling2d (AveragePooling2D)	(None, 1, 12, 40)	0
activation_1 (Activation)	(None, 1, 12, 40)	0
dropout (Dropout)	(None, 1, 12, 40)	0
flatten (Flatten)	(None, 480)	0
dense (Dense)	(None, 2)	962
activation_2 (Activation)	(None, 2)	0

Total params: 33682 (131.57 KB)
Trainable params: 33602 (131.26 KB)
Non-trainable params: 80 (320.00 Byte)

(c)

Figure 7. Summary of the CNN architectures in development environment: (a) EE-GNet; (b) DeepConvNet; and (c) ShallowConvNet.

2.7. Model Assessment

Each architecture was validated using stratified 10-fold cross-validation. For EEGNET and DeepConvNet, folds were generated using the weights obtained in the pretraining step, generating one model per fold for each architecture. The test data were used in each model to generate metrics per fold.

For the ShallowConvNet architecture, cross-validation was performed by loading the weights generated in the pretraining step and performing training and testing per fold in a single step without exporting the generated models in each fold.

The data used for the assessment were the augmented training and testing sets from the previous steps. The models generated per fold were used to obtain the overall accuracy of the model, the receiver operating characteristic (ROC) curve, and metrics derived from their corresponding confusion matrices (precision, recall, and specificity).

 doi:10.20944/preprints202407.1239.v1

3. Results

3.1. Pretraining Results

The performance of the pretrained model for each architecture was assessed in terms of the overall accuracy and loss function during training, and the resulting plots are displayed in Figure 8. Table 1 summarizes the results for the pretraining including the metrics from the confusion matrix and ROC curve for each architecture. Figure 9 shows the confusion matrix for the prediction of each model based on the test data. The ROC curves for the pretrained models are shown in Figure 10.

The EEGNet model provides the highest overall accuracy among the three architectures; however, the DeepConvNet and ShallowConvNet models outperformed EEGNet in terms of recall for the unknown class. From the confusion matrix, a higher recall index in the unknown class is achieved in these models to the detriment of specificity, making them less reliable than EEGNet for predicting trials to belong to the known class. In addition, despite not having significantly higher ratings, EEGNet provided less susceptibility to overfitting and more consistent results for both classes. In addition, the areas under the curves (AUCs) suggest that the three models provide similar degrees of separation, which are slightly higher than random guessing, EEGNet being the architecture with the greatest AUC.

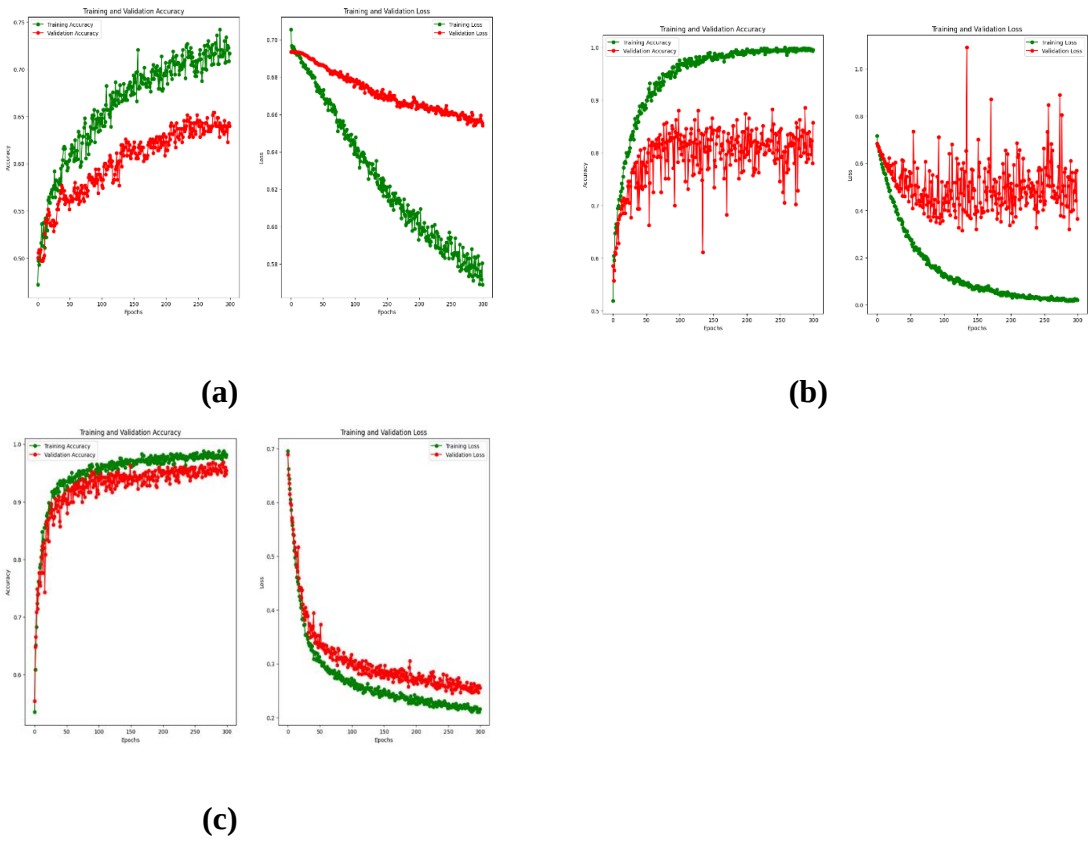


Figure 8. Accuracy and loss function plots for pretrained models: (a) EEGNet; (b) DeepConvNet; and (c) ShallowConvNet. The early stop checkpoint allowed storage of the weights that provided the minimum loss function result and maintained them at the end of the training to avoid overfitting.

Table 1. Summary of metrics for pretrained models.

Model.	**Overall Accuracy**	**ROC AOC**	**Precision**	**Recall**	**Specificity**
EEGNET	0.660	0.670	0.673	0.623	0.697

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 July 2024 doi:10.20944/preprints202407.1239.v1

DeepConvNet	0.550	0.640	0.517	0.869	0.189
ShallowConv Net	0.610	0.640	0.570	0.720	0.620

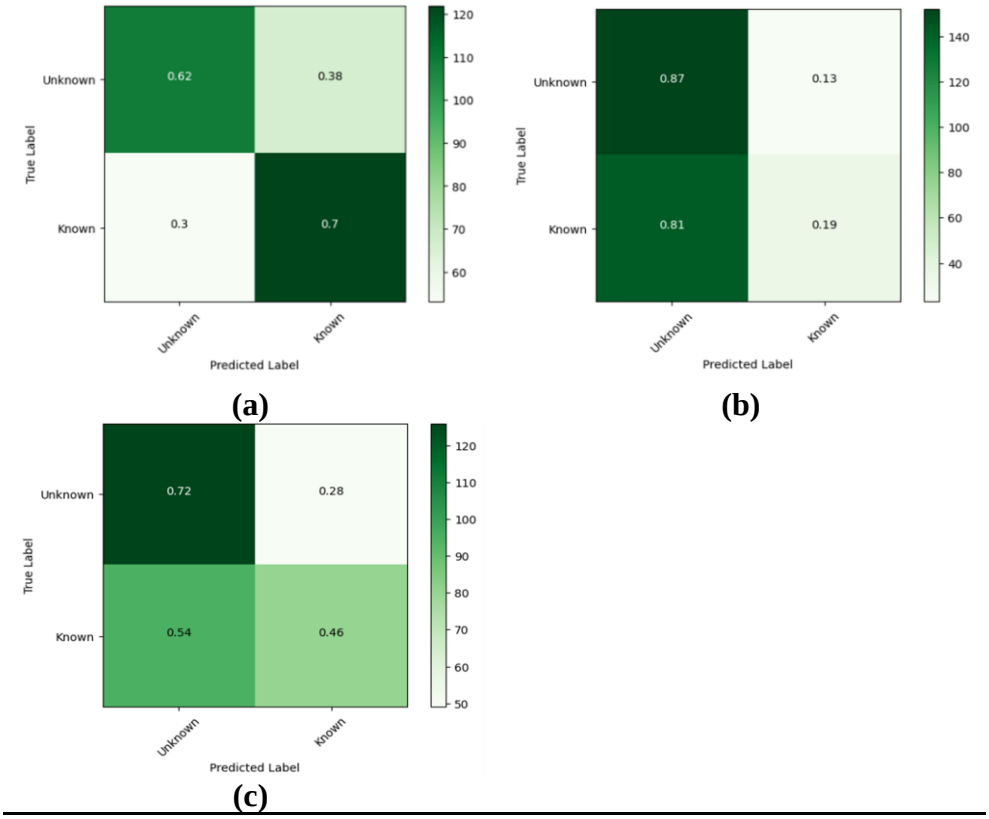
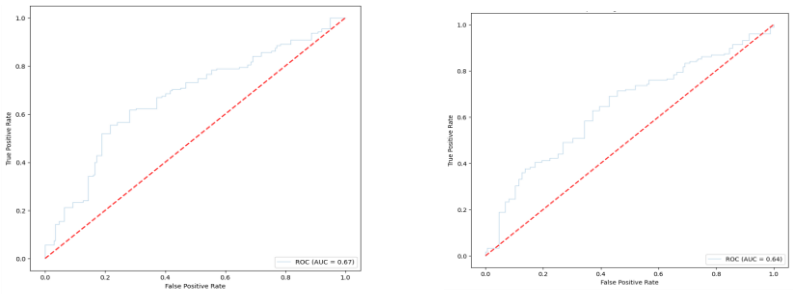


Figure 9. Confusion matrices for pretrained models on test data: (a) EEGNet; (b) DeepConvNet; and (c) ShallowConvNet. The false positive rate for DeepConvNet and ShallowConvNet is significantly higher than that for EEGNet.

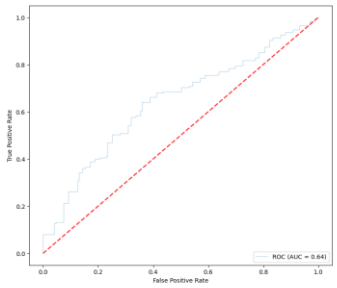
3.2. Cross Validation Results

To assess the general performance of each CNN architecture, the metrics derived from the ROC curves and the confusion matrix metrics calculated for each fold are summarized in Table 2. Figure 11 shows the corresponding ROC curves for each model architecture in each of the folds.



(a) **(b)**

 doi:10.20944/preprints202407.1239.v1



(c)

Figure 10. ROC curves for pretrained models: (a) EEGNet; (b) DeepConvNet; and (c) ShallowConvNet.

Table 2. Metrics of ROC and confusion matrix per fold.

Model	Fold	Overall Accuracy	ROC-AOC	Precision	Recall	Specificity
EEGNET	0	0.68	0.67	0.60	0.64	0.57
	1	0.65	0.65	0.59	0.65	0.55
	2	0.72	0.65	0.61	0.63	0.61
	3	0.77	0.66	0.63	0.6	0.64
	4	0.72	0.66	0.61	0.55	0.65
	5	0.72	0.66	0.59	0.66	0.54
	6	0.76	0.66	0.62	0.59	0.63
	7	0.71	0.66	0.59	0.6	0.59
	8	0.77	0.66	0.59	0.66	0.54
	9	0.65	0.66	0.62	0.62	0.62
	Mean	**0.71**	**0.66**	**0.60**	**0.62**	**0.59**
	SD	**0.04**	**0.01**	**0.01**	**0.03**	**0.04**
DeepConvNet	0	0.96	0.64	0.54	0.9	0.22
	1	0.97	0.63	0.52	0.87	0.19
	2	0.98	0.62	0.51	0.89	0.16
	3	0.96	0.63	0.52	0.86	0.22
	4	0.99	0.64	0.52	0.9	0.18
	5	0.97	0.64	0.55	0.79	0.34
	6	0.99	0.65	0.52	0.88	0.19
	7	0.99	0.64	0.52	0.87	0.19
	8	0.94	0.64	0.53	0.87	0.22
	9	0.96	0.62	0.53	0.83	0.25
	Mean	**0.97**	**0.63**	**0.53**	**0.87**	**0.22**
	SD	**0.02**	**0.01**	**0.01**	**0.03**	**0.05**
ShallowConvNet	0	0.94	0.64	0.58	0.76	0.46
	1	0.9	0.65	0.58	0.77	0.44
	2	0.96	0.64	0.54	0.78	0.34
	3	0.92	0.61	0.54	0.82	0.30
	4	0.97	0.61	0.55	0.79	0.37
	5	0.96	0.63	0.57	0.75	0.43
	6	0.94	0.62	0.55	0.79	0.34
	7	0.91	0.64	0.58	0.77	0.44
	8	0.96	0.63	0.56	0.78	0.39
	9	0.93	0.64	0.56	0.78	0.38
	Mean	**0.94**	**0.63**	**0.56**	**0.78**	**0.39**
	SD	**0.02**	**0.01**	**0.02**	**0.02**	**0.05**

 doi:10.20944/preprints202407.1239.v1

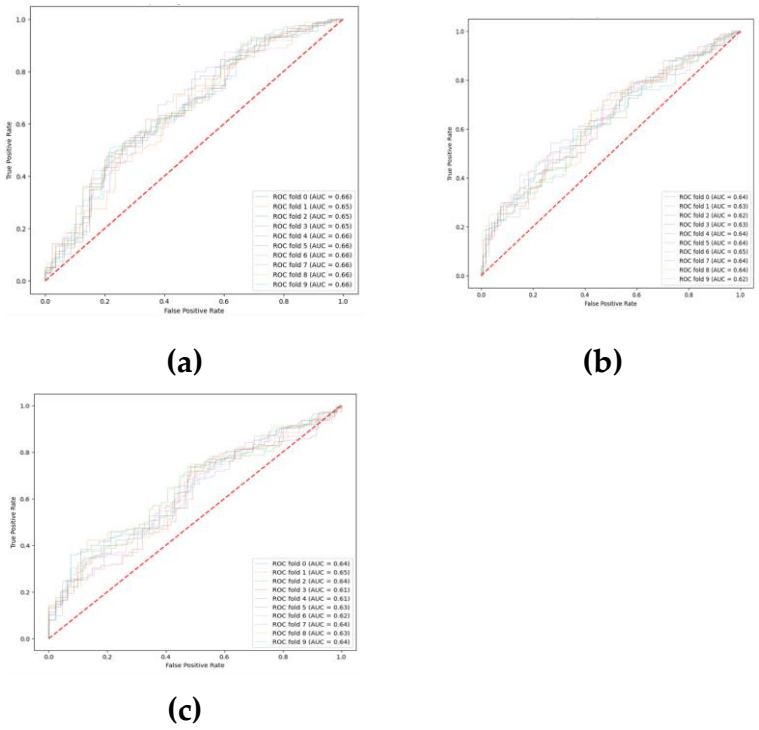


Figure 11. ROC curves for each model during cross-validation: (a) EEGNet; (b) DeepConvNet; and (c) ShallowConvNet. Results are consistent with those obtained from the pretrained models.

The metrics obtained from the cross-validation provide results that are consistent with those obtained from the pretrained models. As summarized in Table 2, despite EEGNET not providing the best overall accuracy, it still outperformed DeepConvNet and ShallowConvNet in terms of specificity and displayed a slightly greater AUC.

4. Discussion

Based on the results obtained from the testing and validation of the different CNN models used in the current study, the reliability of the automatic classification of an encountered item as known or unknown depends on different factors, such as the amount of available data, architecture of the model, and balancing of the data in different classes. Based on the obtained metrics, we conclude that the EEGNet architecture provides a better understanding of the underlying neural mechanisms associated with explicit memory and familiarity than DeepConvNet or ShallowConvNet. Owing to its architecture and the different parameters that can be customized to define the model, EEGNet can potentially better address the most significant features of EEG signals, allowing us to understand familiarity and long-term memory from an electrophysiological perspective. Although DeepConvNet and ShallowConvNet provide higher accuracy and recall indices, they fail to provide a reliable true negative rate compared with EEGNet. Because the problem we are addressing in this study assumes that both classes (known and unknown) are equally important, we consider that a predictive tool based on EEGNet would be more reliable in a real-life context, in which a digital tool is used to assist the study of a set of items. One of the most important challenges that must be overcome when implementing DL-based solutions is the amount of data available for training. Using EEG data to train models is limited by the reduced number of freely available datasets compared with other types of data used for training models, such as images, and the specificity of the collected data regarding the paradigm and purpose, owing to the different responses associated with multiple neural processes. Nevertheless, the models used in this study can still perform predictions with acceptable accuracy, which is consistent with the results obtained in other studies that use CNNs to perform classification based on EEG signals in different contexts and using a rather limited amount

 doi:10.20944/preprints202407.1239.v1

of training data. This could be observed in the AUCs, which are similar in all three models and provide a prediction rate slightly above chance. The problem of the limited amount of training data was partially overcome by oversampling the existing data, allowing a balance of the number of trials belonging to both classes, and hence improving the recall, precision, and specificity indices. However, we believe that the use of real human-balanced data will contribute to increasing the overall performance of the model and provide a more reliable prediction.

5. Conclusion

In this study, we proposed a method to assess familiarity and new/old effects using a DL approach as a surrogate to the traditional grand average method used to identify such phenomena. To the best of our knowledge, this is the first attempt to assess the performance of architectures tested in other paradigms that involve the use of EEG and ERP in an association memory task to predict whether a study item has been previously learned. DL approaches are a promising solution for performing a single-trial analysis of ERP data using classification results and features learned through convolutional layers. The reduced amount of data available for training and the difficulty in obtaining balanced amounts of trials for different classes influence the performance of the models. Future studies on this topic will focus on improving the accuracy of the model by collecting greater amounts of training data and customizing and combining the architectures of different models to obtain more accurate prediction outputs. We expect that the use of CNN models will allow the creation of online tools capable of recording physiological activity in real time and assess the learning process based on the neural response of students.

Author Contributions: Conceptualization: K.H.; Experiment design: J.D., R.M.; Programming: J.D.; Data Analysis: J.D.; Writing the Paper: J.D., R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JST CREST, grant number JPMJCR18A4 and JST Moonshot R&D, grant number JPMJMS2293-04.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of the University of Tokyo (subject code 246-24).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data collected from the experiment are available upon request to the corresponding author. The code containing the experimental task can be accessed at https://github.com/joardemu85/JsPsych_Tasks/tree/main/04_Familiarity_Flags. The code used for analysis can be accessed at https://github.com/joardemu85/AMT_Code.

Acknowledgments: We wish to thank Kashiwakura S., Ling R., Tian Z., Hui Y., and Shimizu K., for their support during our experiment.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Schacter, D. Implicit memory: History and current status. *J. Exp. Psychol. Learn. Mem. Cogn.* **1987**, 13(3), 501–518. https://doi.org/10.1037/0278-7393.13.3.501
2. Cherry, K. (n.d.). Implicit Memory vs. Explicit Memory How the different types of long-term memory work. Available online: https://www.verywellmind.com/implicit-and-explicit-memory-2795346 (Accessed on 24.05.2024)
3. Yonelinas, A. P. The nature of recollection and familiarity: A review of 30 years of research. *J. Mem. Lang.* **2002**, 46(3), 441–517. https://doi.org/10.1006/JMLA.2002.2864
4. Paller, K. A.; Voss, J. L.; Boehm, S. G. Validating neural correlates of familiarity. *Trends Cogn. Sci* **2007**, 11(6), 243–250. https://doi.org/10.1016/j.tics.2007.04.002
5. Wagner, A. D.; Koutstaal, W.; Schacter, D. L. When encoding yields remembering: insights from event-related neuroimaging. Philosophical Transactions of the Royal Society of London. Series B: *Biol. Sci.* **1999**, 354(1387), 1307–1324. https://doi.org/10.1098/rstb.1999.0481

6. Friedman, D.; & Johnson, R. Event-related potential (ERP) studies of memory encoding and retrieval: a selective review. *Microsc. Res. Tech.* **2000**, 51(1), 6–28. https://doi.org/10.1002/1097-0029(20001001)51:1<6::AID-JEMT2>3.0.CO;2-R
7. Curran, T. Effects of attention and confidence on the hypothesized ERP correlates of recollection and familiarity. *Neuropsychologia* **2004**, 42(8), 1088–1106. https://doi.org/10.1016/j.neuropsychologia.2003.12.011
8. Diana, R. A.; van den Boom, W.; Yonelinas, A. P.; Ranganath, C. ERP correlates of source memory: Unitized source information increases familiarity-based retrieval. *Brain Res.* **2011**, 1367, 278–286. https://doi.org/10.1016/J.BRAINRES.2010.10.030
9. Strózak, P.; Abedzadeh, D.; Curran, T. Separating the FN400 and N400 potentials across recognition memory experiments. *Brain Res* **2016**,1635, 41–60. https://doi.org/10.1016/J.BRAINRES.2016.01.015
10. Leynes, P.; Upadhyay, T. Context dissociations of the FN400 and N400 are evidence for recognition based on relative or absolute familiarity. *Brain Cogn.* **2022**, 162. https://doi.org/10.1016/J.BANDC.2022.105903
11. Addante, R. J.; Ranganath, C.; Yonelinas, A. P. Examining ERP correlates of recognition memory: Evidence of accurate source recognition without recollection. *NeuroImage* **2012**, 62(1), 439–450. https://doi.org/10.1016/j.neuroimage.2012.04.031
12. Dimsdale-Zucker, H. R.; Maciejewska, K.; Kim, K.; Yonelinas, A. P.; Ranganath, C. Individual differences in behavioral and electrophysiological signatures of familiarity- and recollection-based recognition memory. *Neuropsychologia* **2022**, 173. https://doi.org/10.1016/j.neuropsychologia.2022.108287
13. Fukuda, K.; Woodman, G. F. Predicting and Improving Recognition Memory Using Multiple Electrophysiological Signals in Real Time. *Psychol. Sci.* **2015**, 26(7), 1026–1037. https://doi.org/10.1177/0956797615578122
14. Khurana, V.; Kumar, P.; Saini, R.; Roy, P. P. EEG based word familiarity using features and frequency bands combination. *Cogn. Syst. Res.* **2018**, 49, 33–48. https://doi.org/10.1016/j.cogsys.2017.11.003
15. Luck, S. J. An introduction to the event-related potential technique, 2nd ed, The MIT Press USA, 2014
16. Luck, S. J. (n.d.). Virtual ERP Boot Camp: Introduction to ERPs. Available online: https://courses.erpinfo.org/courses/take/Intro-to-ERPs/texts/14727771-chapter-1-overview (Accessed on 15.08.2023)
17. Wilding E.; Ranganath C. Electrophysiological Correlates of Episodic Memory Processes In The Oxford Handbook of Event-Related Potential Components, Kappenman E., Luck S. J. Oxford Handbooks. USA, 2012
18. Bandt, C.; Weymar, M.; Samaga, D.; Hamm, A. O. A simple classification tool for single-trial analysis of ERP components. *Psychophysiology* **2009**, 46(4), 747–757. https://doi.org/10.1111/j.1469-8986.2009.00816.x
19. Stober, S.; Cameron, D. J.; Grahn, J. A. Using Convolutional Neural Networks to Recognize Rhythm Stimuli from Electroencephalography Recordings. In NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems. Canada Date of Conference (08.12.2014)
20. Oh, S. L.; Vicnesh, J.; Ciaccio, E. J.; Yuvaraj, R.; Acharya, U. R. Deep convolutional neural network model for automated diagnosis of Schizophrenia using EEG signals. *Appl. Sci*. **2019**, 9(14). https://doi.org/10.3390/app9142870
21. Borra, D.; Magosso, E. Deep learning-based EEG analysis: investigating P3 ERP components. *J. Integr. Neurosci.* **2021**, 20(4), 791–811. https://doi.org/10.31083/j.jin2004083
22. Komolovaitė, D.; Maskeliūnas, R.; Damaševičius, R. Deep Convolutional Neural Network-Based Visual Stimuli Classification Using Electroencephalography Signals of Healthy and Alzheimer's Disease Subjects. *Life* **2022**, 12(3). https://doi.org/10.3390/life12030374
23. Craik, A.; He, Y.; Contreras-Vidal, J. L. Deep learning for electroencephalogram (EEG) classification tasks: A review. *J. Neural Eng.* **2019** (Vol. 16, Issue 3). Institute of Physics Publishing. https://doi.org/10.1088/1741-2552/ab0ab5
24. Depuydt, E.; Criel, Y.; de Letter, M.; van Mierlo, P. Single-trial ERP Quantification Using Neural Networks. *Brain Topogr.* **2023**. https://doi.org/10.1007/s10548-023-00991-8
25. Alimardani, M.; Kaba, M. Deep Learning for Neuromarketing: Classification of User Preference using EEG Signals. In The 12th Augmented Human International Conference (AH2021), Switzerland, Date of Conference (27.05.2021)
26. Atilla, F.; Alimardani, M. EEG-based Classification of Drivers Attention using Convolutional Neural Network. In Proceedings of the 2021 IEEE International Conference on Human-Machine Systems(ICHMS 2021), Germany, Conference (08.09.2021)
27. jsPsych. (n.d.). Available Online: https://www.jspsych.org/7.3/ (Accessed on 24.02.2023)

28. de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. Behavior Research Methods, 47(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y
29. Delorme, A.; Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci.* Methods **2004**, 134(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009
30. Calbi, M.; Siri, F.; Heimann, K.; Barratt, D.; Gallese, V.; Kolesnikov, A.; Umiltà, M. A. How context influences the interpretation of facial expressions: a source localization high-density EEG study on the "Kuleshov effect." *Sci. Rep.* **2019**, 9(1). https://doi.org/10.1038/s41598-018-37786-y
31. Herzmann, G.; Curran, T. Experts' memory: An ERP study of perceptual expertise effects on encoding and recognition. *Mem. Cogn* **2011**, 39(3), 412–432. https://doi.org/10.3758/s13421-010-0036-1
32. Bailey, K.; Chapman, P. When can we choose to forget? An ERP study into item-method directed forgetting of emotional words. *Brain Cogn.* **2012**, 78(2), 133–147. https://doi.org/10.1016/J.BANDC.2011.11.004
33. Lopez-Calderon, J.; Luck, S. J. ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Front. Hum. Neurosci*. **2014**, 8(1 APR). https://doi.org/10.3389/fnhum.2014.00213
34. Luck, S. J.; Stewart, A. X.; Simmons, A. M.; Rhemtulla, M. Standardized measurement error: A universal metric of data quality for averaged event-related potentials. *Psychophysiology* **2021**, 58(6). https://doi.org/10.1111/psyp.13793
35. Schirrmeister, R. T.; Springenberg, J. T.; Fiederer, L. D. J.; Glasstetter, M.; Eggensperger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; Ball, T. Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* **2017**, 38(11), 5391–5420. https://doi.org/10.1002/hbm.23730
36. Lawhern, V. J.; Solon, A. J.; Waytowich, N. R.; Gordon, S. M.; Hung, C. P.; Lance, B. J. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *J. Neural Eng.* **2018**, 15(5). https://doi.org/10.1088/1741-2552/aace8c
37. Lawhern, V. (n.d.). Army Research Laboratory (ARL) EEGModels. Available online: https://github.com/vlawhern/arl-eegmodels (Accessed on 24.02.2024)