frontiers in Neuroscience

ORIGINAL RESEARCH
published: 03 February 2021
doi: 10.3389/fnins.2021.626277

Cross-Subject EEG-Based Emotion Recognition Through Neural Networks With Stratified Normalization

Javier Fdez, Nicholas Guttenberg, Olaf Witkowski and Antoine Pasquali*

Cross Labs, Cross Compass Ltd., Tokyo, Japan

Due to a large number of potential applications, a good deal of effort has been recently made toward creating machine learning models that can recognize evoked emotions from one's physiological recordings. In particular, researchers are investigating the use of EEG as a low-cost, non-invasive method. However, the poor homogeneity of the EEG activity across participants hinders the implementation of such a system by a time-consuming calibration stage. In this study, we introduce a new participant-based feature normalization method, named *stratified normalization*, for training deep neural networks in the task of cross-subject emotion classification from EEG signals. The new method is able to subtract inter-participant variability while maintaining the emotion information in the data. We carried out our analysis on the SEED dataset, which contains 62-channel EEG recordings collected from 15 participants watching film clips. Results demonstrate that networks trained with stratified normalization significantly outperformed standard training with batch normalization. In addition, the highest model performance was achieved when extracting EEG features with the multitaper method, reaching a classification accuracy of 91.6% for two emotion categories (positive and negative) and 79.6% for three (also neutral). This analysis provides us with great insight into the potential benefits that stratified normalization can have when developing any cross-subject model based on EEG.

Keywords: deep learning, feature normalization, stratified normalization, SEED dataset, EEG, cross-subject, emotion recognition, affective computing

OPEN ACCESS

Edited by:
Mohammad Khosravi,
Persian Gulf University, Iran

Reviewed by:
Pooya Tavallali,
University of California, Merced,
United States
Mohammad Kazem Moghimi,
University of Sistan and Baluchestan,
Iran

****Correspondence:***
Javier Fdez
javier@crosslabs.org

Specialty section:
This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience

Received: *05 November 2020*
Accepted: *05 January 2021*
Published: *03 February 2021*

Citation:
Fdez J, Guttenberg N, Witkowski O and Pasquali A (2021) Cross-Subject EEG-Based Emotion Recognition Through Neural Networks With Stratified Normalization. Front. Neurosci. 15:626277. doi: 10.3389/fnins.2021.626277

1. INTRODUCTION

Emotion recognition has gained great attraction due to its large number of potential applications in fields such as human-computer interaction (Brave and Nass, 2009), interactive storytelling (Fels et al., 2011), and mood disorders (El Keshky, 2018). Specifically, researchers are exploiting emotion recognition via EEG signals due to its advantages compared to other low-cost, non-invasive methods such as electromyogram (EMG) and electrocardiography (ECG), whose current limitations restrain them to be used mainly in multimodal emotion recognition (Dzedzickis et al., 2020), or to facial expression and speech emotion recognition methods, which are susceptible to cognitive bias such as social desirability bias (Gery et al., 2009; Heuer et al., 2007).

However, the main bottleneck in the development of models trained with EEG signals is the poor homogeneity of between-sessions data and between-participants data, which, interestingly, is not apparent in the literature in the context of emotion recognition from facial expressions or other physiological data (Cimtay and Ekmekcioglu, 2020). In order to solve this problem with EEG, current methods rely on participant-dependent models tuned with tedious and time-consuming calibration sessions implemented before each experiment.

In the past years, a significant effort has been made in building participant-independent models that eliminate the need for calibration sessions. Specifically, the primary focus of these models is to find common features across participants using algorithms which usually regard variance among individuals as mere statistical noise (Shu et al., 2018). One example is the study by Li et al. (2020), where researchers used an unsupervised deep generative model to capture the emotion-related information between participants. Another example is from Yin et al. (2017), who present an EEG feature selection approach to determine a set of the most robust EEG indicators with stable geometrical distribution across a group of participants. In another study by Li et al. (2018), researchers extracted nine types of time-frequency domain features and nine types of dynamical system features and studied the importance of all those features across different channels, brain regions, rhythms, and features types. Last but not least, the study by Song et al. (2020) proposes a graph to model the multichannel EEG features and then perform EEG emotion classification based on this model.

Since the average classification accuracy by selecting robust features is still lower than the participant-dependent models (Shu et al., 2018), researchers are also investigating other approaches such as functional brain connectivity patterns, domain adaptation, or hybrid methods. An example of cross-subject functional brain connectivity investigation is from Cao et al. (2020), who studied the key information flow of the different parts of the brain with minimum spanning trees (MST). About domain adaption, the study by Chai et al. (2016) presents several unsupervised domain adaptation techniques based on autoencoders for non-stationary EEG-based emotion recognition. Furthermore, Cimtay and Ekmekcioglu (2020) analyzes the use of pre-trained convolutional neural network (CNN) architectures to improve the feature extraction and inherent exploitation of the domain adaptation. Lastly, the study by Yang F. et al. (2019) gives an example of a hybrid method for cross-subject emotion recognition by extracting multiple features for the formation of high-dimensional feature space.

Another approach in transfer learning classification tasks, not only relevant in neuroscience but also in image processing, machine learning or pattern recognition, is data normalization. According to Milligan and Cooper (1988), data normalization not only simplifies the numerical calculation, which may help to speed up the learning process during the backpropagation, but also allows the data to have similar dynamic range.

In the context of neuroscience, researchers have investigated and compared different normalization methods for both participant-dependent and participant-independent models. For example, the study by Yousif et al. (2020) compared the performance of three different types of feature normalization—Z-score, min-max, and decimal-scaling normalization—vs. non-normalization in EEG signal-based emotion classification, achieving optimal performance with the Z-score normalization method. Issa and Shedeed (2016) also compared normalized and non-normalized sets of data using four types of feature extraction methods, concluding as well that the normalization procedure enhanced the performance and increased the classification accuracy. Besides, Logesparan et al. (2011) assessed five previously reported normalization techniques—mean memory, standard deviation memory, peak detector, signal range method, and median decaying memory—to correct the amplitude differences in recorded signals between different patients. They discovered that only the last method improved accuracy, which indicates the importance of selecting an appropriate normalization method.

Regarding the importance of normalization in cross-subject classification, the studies by Koelstra et al. (2012) and Jatupaiboon et al. (2013) already gave the first insights into the advantages of this approach after applying participant-based data normalization to reduce the inter-participant variability. They observed that the distance between the clusters, where each cluster corresponds to one participant since the contribution of the participant identification information is higher than the effect of the emotion (Arevalillo-Herráez et al., 2019), reduces when normalizing the data for each participant independently. Later on, the work of Arevalillo-Herráez et al. (2019) exploited this result and proposed a nonlinear data transformation using the median for each feature and participant that seamlessly integrated individual traits into an inter-participant approach. After applying the proposed transformation, they trained a classifier and compared their results with a standard Z-score standardization. Despite proving that their method was able to reduce the magnitude of this component when using PSD features and showing that their results overpassed the Z-score standardization's performance, they noticed in their study that it is necessary to find new normalization methods since the removal of the subject-dependent component in the signal is indeed feature and problem dependent.

In all the mentioned articles, to the best of the author's findings, the researchers' effort has been in reducing the inter-participant variability using data normalization in the pre-processing or before the training stage. However, they have not considered the large benefits that data normalization may have when included within the training process. This is, in fact, a common practice in deep learning with the well-known normalization method called *batch normalization* (Ioffe and Szegedy, 2015), which was first introduced to address the problem of internal covariate shift or an unwanted drift in the distribution of neuron's activations resulting from the learning process. Nevertheless, this method should not in principle reduce the inter-participant variability substantially since it normalizes per batch of data, not distinguishing across participants as Arevalillo-Herráez et al. (2019) mentioned in his work.

In recent years, the rapid progress of machine learning has brought new normalization methods applied in different machine learning fields such as layer normalization (Ba et al., 2016), commonly used in recurrent neural networks (RNN), and instance normalization (Ulyanov et al., 2017), applied in style transfer so that output stylized images do not depend on the

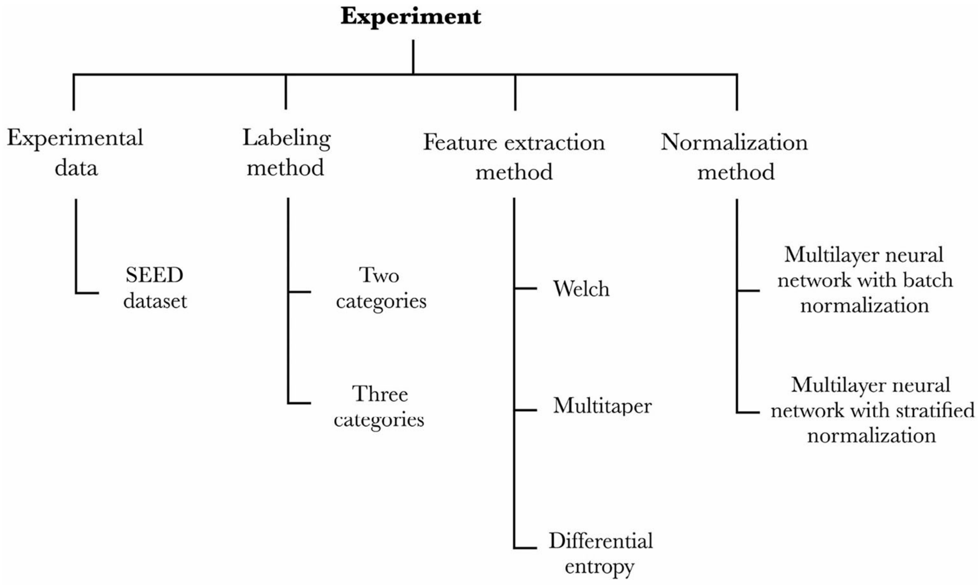

FIGURE 1 | Our experimental setup captures the effects that the experimental data, labeling method, feature extraction method, and normalization method have onto the emotion recognition accuracy of the models.

contrast of the input content image. Interestingly, we observed that the latter method resembled our approach where, instead of removing the contrast of the input image, we were aiming to remove the participant identification information contained in the data.

Therefore, we designed and assessed a new participant-based feature normalization approach, named *stratified normalization*, that normalizes the data per feature and participant within the layers of the neural network classifier. Compared with instance normalization, which can be only used with images and does not normalize across the batch, our method uses the participant labels as additional information to normalize the features and can be extended to any type of data. To deepen the evaluation and assess whether our method reduces the inter-participant variability, we compared our results with standard batch normalization using different labeling and feature extraction methods.

To encourage further research on these topics, we have made the source code of this work freely accessible to all[1].

2. MATERIALS AND METHODS

We designed an experiment to capture the effects of one control variable—the experimental data—and three independent variables—the labeling method, the feature extraction method, and the normalization method—onto the dependent variable that we aim to assess—the cross-subject emotion recognition accuracy of the trained models. **Figure 1** details the variables and their conditions for the experiment. The four sections of this chapter detail each of the experiment's control and independent variables, respectively.

Our methodology unfolds as follows. Firstly, we picked an EEG dataset and a preprocessing stage of the data so as to feed our models (c.f., section 2.1). Secondly, we prepared two strains of models to perform either binary or ternary classification (c.f., section 2.2). Subsequently, we defined models that would extract the features according to the three conditions of the feature extraction method variable (c.f., section 2.3). Lastly, we implemented the two conditions of the normalization method variable (c.f., section 2.4) and assessed each model ($2 \times 3 \times 2 = 12$ models in total) through a leave-one-out cross-validation over all the participants from the EEG dataset.

2.1. Experimental Data

Despite the large number of potential applications of emotion recognition from EEG signals, to the best of our knowledge, MAHNOB-HCI (Soleymani et al., 2012), SEED (Duan et al., 2013; Zheng and Lu, 2015), SEED IV (Zheng et al., 2019), and DEAP (Koelstra et al., 2012) are the only four publicly available emotional EEG datasets on the topic. From those four datasets, the most studied ones in cross-subject emotion classification are the DEAP and SEED datasets (e.g., Li et al., 2020; Cimtay and Ekmekcioglu, 2020, from which we decided to focus on the SEED dataset because of the following reasons. Firstly, the collected data is well-balanced between sessions and participants, which simplifies the assessment of the classification results. Secondly, as Li et al. (2018) mentioned in his study, the performance on DEAP is significantly inferior to that on SEED, which may be due to the relatively low quality of the data and the emotional ratings of trials. This led them to conduct further evaluation only in the SEED dataset. Thirdly, neural networks

[1] https://github.com/javiferfer/Cross-subject-EEG-emotion-recognition-through-NN

easily overfit with the DEAP dataset, which would increase the complexity of the neural network models since it would be necessary to include data augmentation, early stopping, higher dropout, and/or regularization for each model (e.g., Cimtay and Ekmekcioglu, 2020; Yang H. Y. et al., 2019). Lastly, the literature offers many reports on cross-subject emotion recognition models using the SEED dataset, which permits the comparison with other papers.

The SEED dataset contains 62-channel EEG data collected from 15 participants, who carried out three sessions over the same 15 film clips. An emotional rating was previously assigned to each film clip and obtained by averaging the ratings of 20 participants who were asked to indicate one keyword (positive, neutral, or negative) after watching them. In the subsequent experiment with EEG recordings, the clips used across sessions and participants at each trial shared the same pseudo-random order of ratings to balance evoked emotions throughout the EEG recordings smoothly.

In addition to the raw EEG data, the SEED dataset contains a preprocessed version of the signals which consisted of downsampling to 200 Hz and noise filtering with a bandpass filter of 0.5–70 Hz. Since the downsampling reduces the high dimensionality, and the noise filtering increases the signal-to-noise ratio (Bigdely-Shamlo et al., 2015), we have chosen to use this preprocessed data for our analysis.

2.2. Labeling Method

Our first independent variable indicated the number of emotional classes: two categories (positive and negative) and three categories (positive, neutral, and negative). The two approaches can be conveniently compared due to the large number of articles that reported their results with binary classification (e.g., Li et al., 2018; Yang F. et al., 2019; Li et al., 2020; Cimtay and Ekmekcioglu, 2020) or ternary classification (e.g., Chai et al., 2017; Zhang et al., 2019; Lan et al., 2019; Cimtay and Ekmekcioglu, 2020).

2.3. Feature Extraction Method

The second independent variable concerns the feature extraction method, which refers to either the (1) Welch, (2) multitaper, or (3) Differential Entropy (DE) method in our study. The first two methods were selected since they belong to the Power Spectral Density (PSD) category, which, according to Craik et al. (2019), is a typical approach when training deep neural networks for EEG classification tasks. Our selection of DE was based on the high accuracy reported in some EEG emotion recognition studies such as by Duan et al. (2013) and by Chen et al. (2019).

For all three methods, the total number of extracted features for each trial is 248 (62 channels × 4 band frequencies). The four band powers for each EEG signal correspond to the theta rhythm (4–7 Hz), alpha rhythm (8–13 Hz), beta rhythm (14–30 Hz), and gamma rhythm (31–50 Hz). The delta rhythm (0.5–4 Hz) was excluded as it is traditionally associated with sleep stages (De Andrés et al., 2011) and therefore assumed to be less relevant to our study.

2.3.1. Welch's Method

Welch's method (Welch, 1975) is an approach to estimate the power of a signal at different frequencies. It is carried out by averaging consecutive periodograms of small time-windows over the signal. To encompass at least two full cycles of the lowest frequency of interest (4 Hz), the duration of the time-windows was set at 0.5 s, with an overlap of 0.25 s between each consecutive window. To smooth the discretization process, each window was filtered with a Hann function. The band frequencies were thereafter extracted from the PSD by implementing Simpson's rule, which approximates integrals using quadratics.

2.3.2. Multitaper Method

Multitaper method is an alternative to Welch's method, which still produces high variance for the direct spectral estimation (Mansouri and Castillo-Guerra, 2019). This method reduces the variance of the spectral estimation by using multiple time-domain windows rather than a single-domain window. As well as Welch's method, the band frequencies were extracted by implementing Simpson's rule over the PSD.

2.3.3. Differential Entropy Method

The DE is used to measure the complexity of a continuous random variable (Duan et al., 2013). Its calculation formula can be expressed as,

$$h(X) = -\int_{X} f(x)log(f(x))dx \tag{1}$$

where X is a random variable and f(x) is the probability density function of X. When the time series X obeys the Gaussian distribution $N(\mu, \sigma^2)$, its differential entropy can be defined as,

$$\begin{aligned} h(X) &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} log(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}})dx \\ &= \frac{1}{2} log(2\pi e\sigma^2) \end{aligned} \tag{2}$$

For a fixed-length EEG sequence, we can approximate the EEG data to a Gaussian distribution, constructing features using Equation (2).

2.4. Normalization Method

The last independent variable is the normalization method, which indicates the two types of classifiers implemented in this study: multilayer neural network with batch normalization and multilayer neural network with stratified normalization.

The models used to classify emotions are independent of the normalization method, meaning that batch and stratified normalizations apply to the same neural architectures (c.f., **Figure 2**). Both types of classifiers were trained for 100 epochs; the whole training data (15 participants × 3 sessions × 15 trials) was input in one unique batch; the learning rate was set to 0.005 for the first 40 epochs, then decreased to 0.001 for the remaining 60 epochs; the optimizer was the Adam optimizer (Kingma and Ba, 2014); the loss function was the negative log-likelihood loss (Zhu et al., 2020). For further information about the hyperparameters or architecture of the neural network, please refer to the source code.

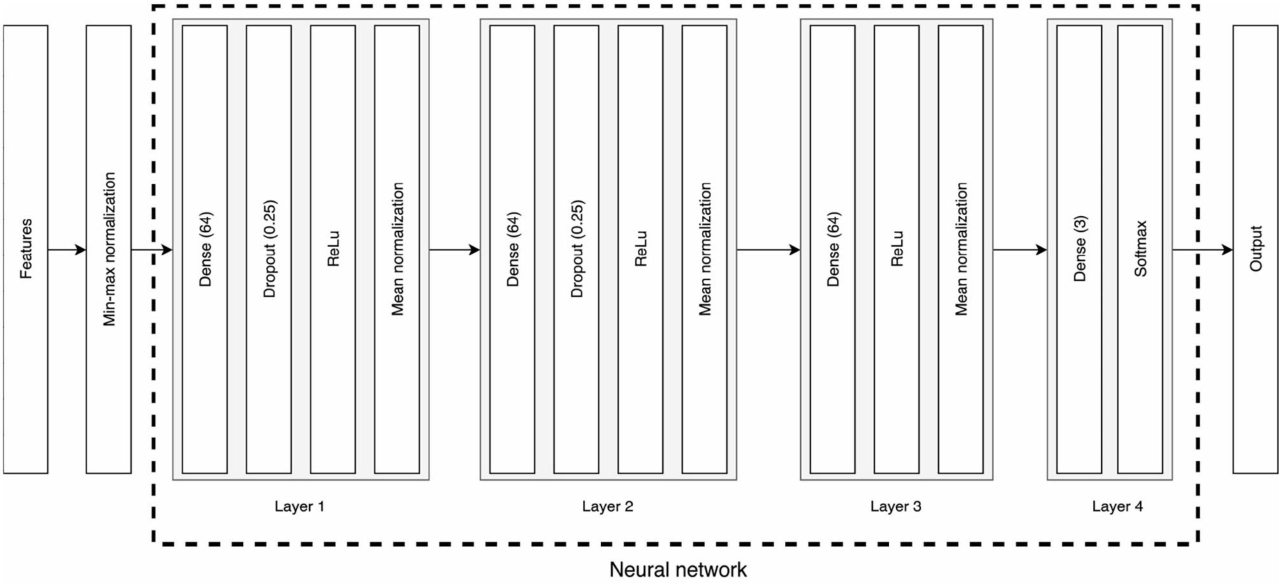

FIGURE 2 | Architecture of the classifiers. Features first go through a min-max normalization of the data before being input to the neural network. The first three layers consist of dense layers with 64 neurons, a dropout for the first two, ReLu activations, and either a batch or a stratified normalization. The last layer is a 3-neuron dense layer that outputs the classification prediction through a Softmax function.

2.4.1. Batch Normalization

Batch normalization was first introduced by Ioffe and Szegedy (2015) in order to address the problem of *internal covariate shift*, an unwanted drift in the distribution of neuron's activations resulting from the learning process. As explained further below, we slightly adapted the method for our purposes. **Figure 3** illustrates our implementation of the batch normalization method.

The extracted features are first min-max normalized according to the following equation:

$$\hat{x}_{ijk} = \frac{x_{ijk} - min(x_k)}{max(x_k) - min(x_k)} \tag{3}$$

where the parameter *i* indicates the number of the participant and session (45 in total), *j* refers to the number of the trial (15 in total), and *k* identifies the number of the feature (248 in total).

The output of this min-max normalization is input to the neural network, which implements mean normalization of the output of each of the first three layers according to Equations (4), (5), and (6).

$$\mu_k = \frac{1}{45 * 15} \sum_{i=1}^{45} \sum_{j=1}^{15} x_{ijk} \tag{4}$$

$$\sigma_k^2 = \frac{1}{45 * 15} \sum_{i=1}^{45} \sum_{j=1}^{15} (x_{ijk} - \mu_k)^2 \tag{5}$$

$$\hat{x}_{ijk} = \frac{x_{ijk} - \mu_k}{\sqrt{\sigma_k^2 + \varepsilon}} \tag{6}$$

These equations correspond to the first three steps of the batch normalization transform described by Ioffe and Szegedy (2015). The fourth step of the algorithm is a scale and shift of the normalized values, where the parameters are learned along with the original model parameters. However, we observed that this step decreased the emotion recognition accuracy, so we decided to exclude it from our analysis.

2.4.2. Stratified Normalization

The stratified normalization consists of a feature normalization per participant and session. **Figure 4** details our implementation of the stratified normalization method.

The extracted features are first min-max normalized according to the following equation:

$$\hat{x}_{ijk} = \frac{x_{ijk} - min(x_{ik})}{max(x_{ik}) - min(x_{ik})} \tag{7}$$

The output of this min-max normalization is input to the neural network, which implements mean normalization at the output of each of the first three layers according to Equations (8), (9), and (10).

$$\mu_{ik} = \frac{1}{15} \sum_{j=1}^{15} x_{ijk} \tag{8}$$

$$\sigma_{ik}^2 = \frac{1}{15} \sum_{j=1}^{15} (x_{ijk} - \mu_{ik})^2 \tag{9}$$

$$\hat{x}_{ijk} = \frac{x_{ijk} - \mu_i k}{\sqrt{\sigma_{ik}^2 + \varepsilon}} \tag{10}$$

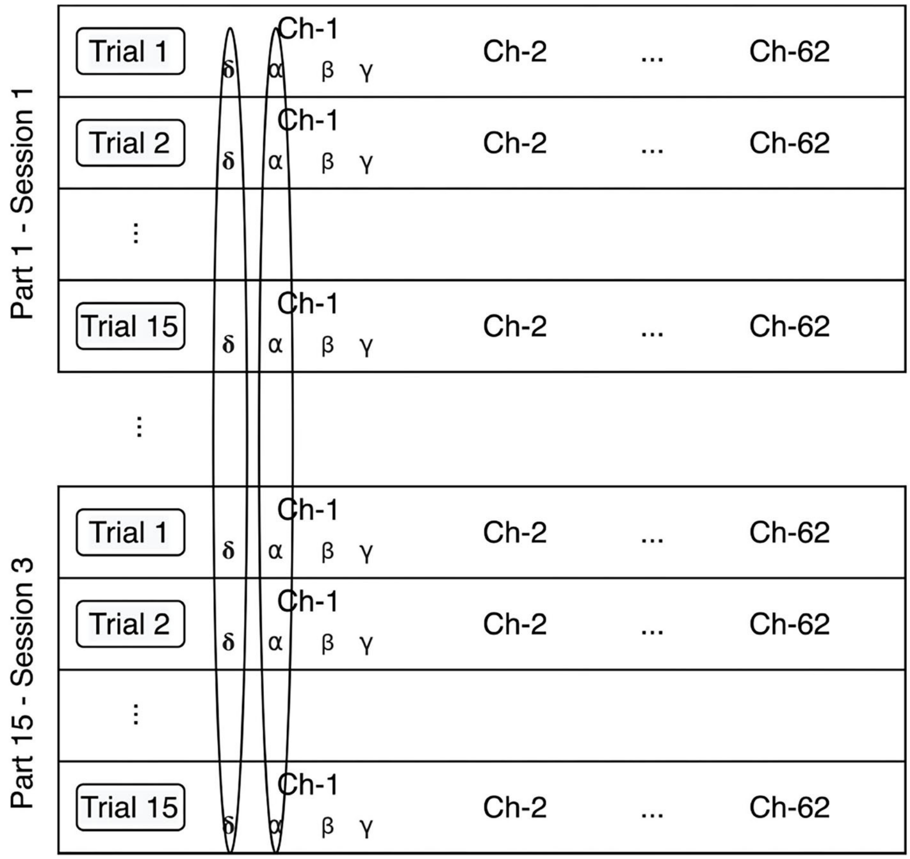

FIGURE 3 | Batch normalization method. The data is normalized per feature, independently of the participant, and session.

3. RESULTS AND DISCUSSION

This section first presents the results of the experiment, then analyzes the between-participant variance and cross-subject emotion recognition in the layers of the neural networks, and finally compares the results of this work with state-of-art literature.

3.1. Overall Evaluation

Figure 5 reports the performance, tested after 100 epochs of training, of the models in each experimental condition.

To evaluate the performance of the models, we ran a three-way ANOVA where the between factors were the (1) labeling, (2) normalization, and (3) feature extraction methods. The statistical results revealed an effect of the labeling method [$F_{(1,\ 168)} = 97.7, p < 0.001, \eta_p^2 = 0.368$], meeting our expectations of an overall better performance in task of binary classification compared to ternary classification. A strong effect of the normalization method was captured as well [$F_{(1,\ 168)} = 33.8$, $p < 0.001$, $\eta_p^2 = 0.17$], confirming that stratified normalization is a more efficient approach than batch normalization in such an experimental context. However, our manipulation of the feature extraction method did not elicit any effect [$F_{(2,\ 168)} = 0.32$, $p = 0.73, \eta_p^2 = 0.004$], meaning that all methods have the potential to perform equally well. No two-way or three-way interactions were captured by our analysis either (Fs < 4).

Despite a lack of effect of the feature extraction method, we have decided to look deeper into the performance of models according to each feature extraction method. Our aim was to allow comparison with the state-of-the-art literature and to further deepen our theoretical interpretation of the results. For binary classification, the methods that performed with the highest accuracy for batch and stratified normalization were DE (M = 0.876, SD = 0.101) and multitaper (M = 0.916, SD = 0.074), respectively. For ternary classification, Welch (M = 0.671, SD = 0.088) and multitaper (M = 0.796, SD = 0.104) were the optimal methods for batch and stratified normalization, respectively. Thus, models using the multitaper feature extraction method in combination with stratified normalization elicit the highest performance both in binary and ternary classification tasks. We will therefore focus on these models for the remaining of our study.

3.2. Descriptive Summary for the Multitaper and Stratified Methods

In this section, we present a descriptive summary of the results obtained when using, as normalization method, the stratified

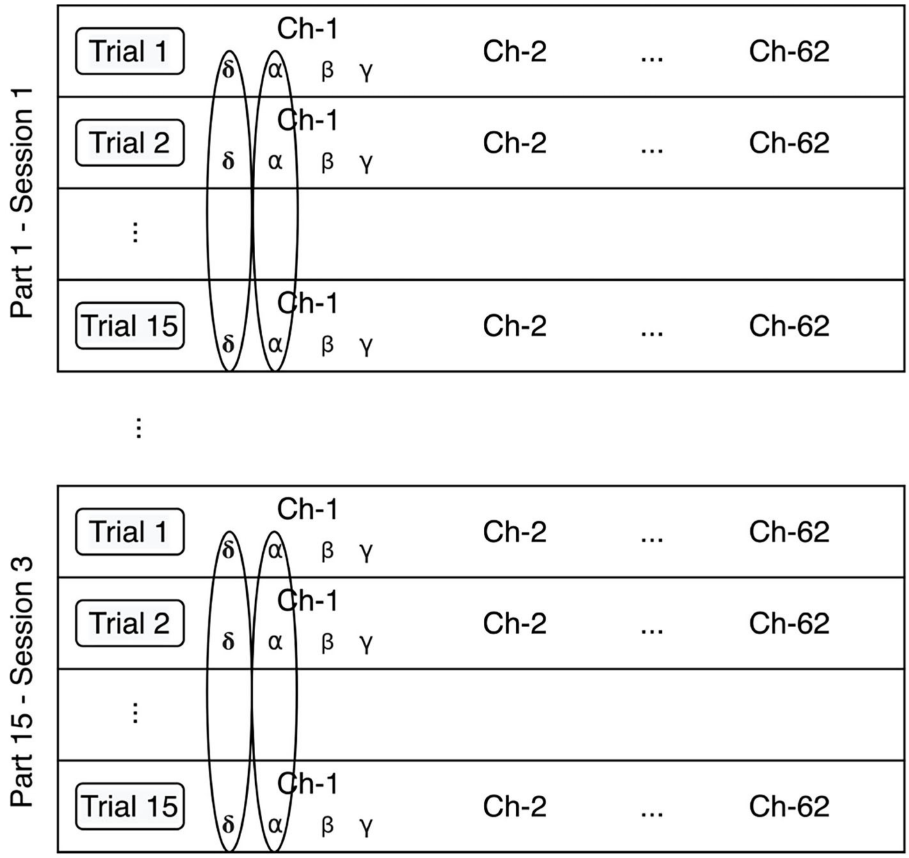

FIGURE 4 | Stratified normalization method. The data is normalized per feature, participant, and session.

normalization and, as feature extraction method, the multitaper method. We selected the former method since it resulted in being statistically significant compared to batch normalization in terms of models' accuracy. About the feature extraction method, the statistical test did not draw any significant result that could allow us to conclude on which method is the most effective. However, we decided to select the multitaper method since it was the feature extraction method with which models performed with the highest accuracy on average, for both binary ($M = 0.916$, SD $= 0.074$) and ternary classification ($M = 0.796$, SD $= 0.104$).

Table 1 indicates the leave-one-out classification accuracies for models based on the multitaper and stratified normalization methods. Each column represents the test accuracy of the models on the untrained data (1 participant out of 15) throughout our 15-fold cross-validation design. The comparison between binary and ternary classification results highlights a moderate correlation between the models' performance on these two tasks (Pearson correlation of 0.343), suggesting that models extract participant identification information with some consistency over tasks (but see section 3.3).

The confusion matrices for the multitaper method when combining the results of the leave-one-out cross-validation of all the participants for two and three categories are shown in **Tables 2**, **3**, respectively. The classification accuracy for positive (92.44%) and negative (90.67%) labels is well-balanced in binary classification. However, for ternary classification, while the classifiers keep an accurate performance for positive labels (90.67%), their accuracy is lowered for negative (75.11%) and neutral (78.34%) labels, indicating that introducing the neutral labels hinders the classification of the negative labels.

Table 4 lists a performance comparison between state-of-the-art models and our proposed method for two categories (positive and negative) and three categories (positive, neutral, and negative).

For binary classification, the last benchmark was reported by Yang F. et al. (2019), who themselves obtained an accuracy of 89.0% by first extracting multiple features for the formation of high-dimensional features and then integrating the significance test/sequential backward selection with the support vector machine for the classification. Our best accuracy for binary classification is 91.6%, which overpasses all reported methods.

For ternary classification, our best accuracy is 79.6%. From the articles listed, our proposed method is overpassed by Zhang et al. (2019), whose model based on convolutional neural network (CNN) and deep domain confusion (DDC) achieved an accuracy of 82.1%, and by Chai et al. (2017), who reached 80.4% by using adaptive subspace feature matching (ASFM). Nevertheless,

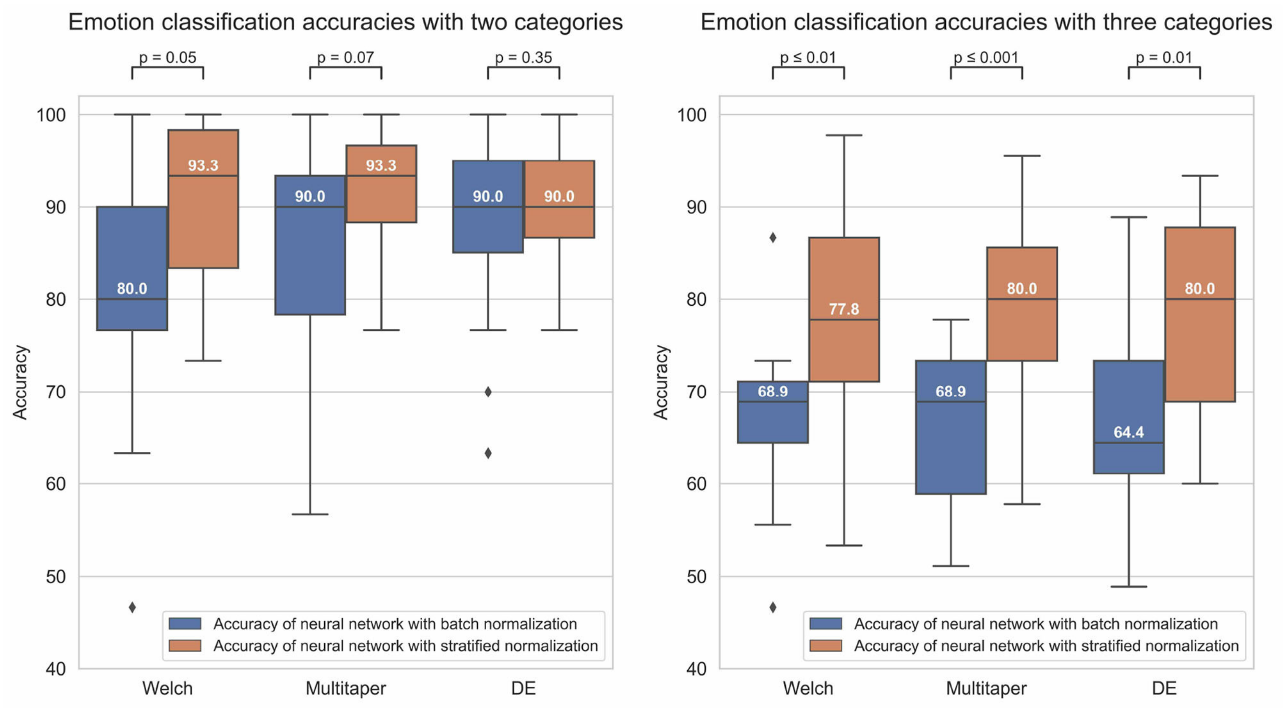

FIGURE 5 | The boxplots of the figure indicate the distribution of the emotion recognition accuracies (y-axis) of the models in each experimental condition, with either (1) a binary classification (left graph) or ternary classification (right graph) labeling method, (2) a Welch, multitaper, or Differential Entropy (DE) feature extraction method (graph's columns), (3) a batch (blue bars) or stratified (orange bars) normalization method. The value inside the boxplots is the median value of the distribution. Besides, the figure also reports the *p*-value of the dependent-samples *t*-test between the two normalization method conditions for each feature extraction method condition.

TABLE 1 | Leave-one-out classification accuracies for two and three categories.

Participant No.	**s01**	**s02**	**s03**	**s04**	**s05**	**s06**	**s07**	**s08**	**s09**	**s10**	**s11**	**s12**	**s13**	**s14**	**s15**
Pos-Neg	76.7	80.0	93.3	86.7	96.7	90.0	93.3	100.0	96.7	96.7	100.0	80.0	93.3	90.0	100.0
Pos-Neu-Neg	77.8	57.8	93.3	82.2	84.4	80.0	80.0	91.1	84.4	71.1	75.6	86.7	62.2	71.1	95.6

Participant numbers (s01–s15) correspond to each test data of the 15-fold cross-validation.

TABLE 2 | Confusion matrix when combining the results of the leave-one-out cross-validation of all the participants for two categories.

		Predicted label		
		Positive	**Negative**	
True label	Positive	208	17	92.44%
	Negative	21	204	90.67%
		90.83%	92.31%	

TABLE 3 | Confusion matrix when combining the results of the leave-one-out cross-validation of all the participants for three categories.

		Predicted label			
		Positive	**Neutral**	**Negative**	
True label	Positive	204	8	13	90.67%
	Neutral	14	170	41	75.56%
	Negative	23	39	163	72.44%
		84.65%	78.34%	75.11%	

compared with our approach, both Zhang et al. (2019) and Chai et al. (2017) used the validation set during the training process, which can increase the cross-subject accuracy.

3.3. Evaluation on Input, Hidden, and Output Layers of the Neural Network

In the previous sections, we evaluated the normalization methods and established that stratified normalization improves the cross-subject emotion recognition accuracy. Nevertheless, an interesting question to be asked is at what depth of the neural network does stratified normalization help increase the accuracy. To answer this question, we further analyzed the emotion recognition accuracy of the models, and their ability to capture and exploit the *brain signature* of each participant—defined as the part of information extracted from the brain signals

TABLE 4 | Performance comparison between this work and the state-of-art literature.

References	SEED performance	
	Acc (Pos-Neg)	Acc (Pos-Neu-Neg)
Chai et al. (2017)	–	80.4
Li et al. (2018)	83.33	–
Zhang et al. (2019)	–	82.1
Lan et al. (2019)	–	72.47
Yang F. et al. (2019)	89.0	–
Li et al. (2020)	85.81	–
Cimtay and Ekmekcioglu (2020)	86.5	78.3
This work	91.6	79.6

that is specific to that participant [also called *subject-related component* by Arevalillo-Herráez et al. (2019)], such that *it can directly inform us on which participant it's been extracted from*. Intuitively, we would expect the emotion recognition accuracy to increase with each layer of the neural network, while the brain signature would fade out due to a decrease of the inter-participant variability with each data normalization.

Therefore, we carried out the evaluation in accordance with the following methodology. For starters, we retrained our models in a three-fold cross-validation design, using the data of 10 participants for training and 5 for testing. For each of the three testing sessions, we recorded the output, or *predicted values*, of each of the model's layers (after normalization for the input, first hidden, second hidden, and third hidden layers, and after softmax for the output layers). We then fed these predicted values to a series of Support Vector Machines (SVM) with RBF kernels (Chang et al., 2010). To do so, we mixed the data of all 5 test participants and ran a five-fold cross-validation per layer on the following tasks. In one case, we used the emotional ratings of the five test participants as labels, thus evaluating the capacity of each layer to contribute to the emotion recognition accuracy of the model. In another case, we used as labels the participant identification numbers (from 1 to 5), here evaluating the amount of brain signature still available at each layer. The classification results are shown in **Figure 6**. We also separated the dataset between two and three emotional categories for running the analyses.

The statistical analysis of the results is carried out in the following subsections. The dependent variable to study is the classification accuracy obtained. The between factors are the (1) labeling method (two categories, three categories), (2) normalization method (batch or stratified normalization), and (3) layer's depth (input layer, first layer, second layer, third layer, output layer).

3.3.1. Emotion Recognition

To evaluate the emotion recognition in the layers of the neural network, we first implemented a three-way ANOVA test. Results indicated no effect of the layer's depth [$F_{(4,\,40)} = 1.49$, $p = 0.22$, $\eta_p^2 = 0.13$]. However, we were able to capture main effects of labeling [$F_{(1,\,40)} = 401.4$, $p < 0.001$, $\eta_p^2 = 0.91$] and of normalization [$F_{(1,\,40)} = 541.0$, $p < 0.001$, $\eta_p^2 = 0.93$] methods. Besides, results indicated that there was an interaction between the labeling method and the normalization method [$F_{(1,\,40)} = 7.38$, $p = 0.010$, $\eta_p^2 = 0.16$]. Hence, we implemented a two-way ANOVA test for each of the two conditions of the labeling method. For both of them, the normalization method was found statistically significant [$F_{(1,\,20)} = 180.5$, $p < 0.001$, $\eta_p^2 = 0.90$ for two categories and $F_{(1,\,20)} = 405.9$, $p < 0.001$, $\eta_p^2 = 0.95$ for three categories].

Although the ANOVA didn't capture any other interaction with layer's depth (all Fs < 2), we also evaluated normalization methods separately at the input and output layers of the neural network. Results of two-way ANOVA tests for the input layer [$F_{(1,\,8)} = 156.8$, $p < 0.001$, $\eta_p^2 = 0.95$] and the output layer [$F_{(1,\,8)} = 114.1$, $p < 0.001$, $\eta_p^2 = 0.93$] both indicated that stratified normalization surpasses batch normalization.

Therefore, (1) the emotion recognition accuracy does not increase significantly along with the layers of the neural network, and (2) the stratified normalization outperforms batch normalization in emotion recognition, as we already concluded above.

3.3.2. Participant Identification

To analyze results in terms of the participant identification accuracy, we first ran a three-way ANOVA test, which pointed out an interaction between the three factors [$F_{(4,\,40)} = 3.44$, $p = 0.017$, $\eta_p^2 = 0.256$], and main effects of the labeling method [$F_{(1,\,40)} = 6.15, p < 0.05, \eta_p^2 = 0.13$], of the normalization method [$F_{(1,\,40)} = 401.1$, $p < 0.001$, $\eta_p^2 = 0.91$] and of the layer's depth [$F_{(4,\,40)} = 57.27, p < 0.001, \eta_p^2 = 0.85$]. As a result, we carried out a two-way ANOVA test for each condition of the normalization method separately.

For the batch normalization method, we found a significant interaction between the labeling method and the layer's depth [$F_{(4,\,20)} = 3.07, p = 0.040, \eta_p^2 = 0.38$]. Hence, we implemented a one-way ANOVA test for each condition of the labeling method, each revealing effects of the layer's depth [$F_{(4,\,10)} = 60.3$s, $p < 0.001$, $\eta_p^2 = 0.96$ for two categories and $F_{(4,\,10)} = 15.7$, $p < 0.001, \eta_p^2 = 0.87$ for three categories]. Specifically, we found out by implementing several dependent-samples t-tests between each pair of consecutive layers that the accuracy only drops significantly between the third and the output layers [$t_{(4)} = 6.54$, $p = 0.003$, d = 7.70 for two categories and $t_{(4)} = 3.72$, $p = 0.043$, d = 10.7 for three categories].

For the stratified normalization method, analyses showed a significant effect of the layer's depth [$F_{(4,\,20)} = 26.9$, $p < 0.001$, $\eta_p^2 = 0.84$]. Subsequently, we implemented a two-way ANOVA test between each pair of consecutive layers and this time, we captured a significant drop in accuracy between the input and the first layer [$F_{(1,\,8)} = 43.6$, $p < 0.001$, $\eta_p^2 = 0.84$]. This result suggests a strong correlation between the use of the stratified normalization method and an early drop in participant identification accuracy.

Lastly, to determine if the normalization methods were statistically different, we computed a two-way ANOVA test for

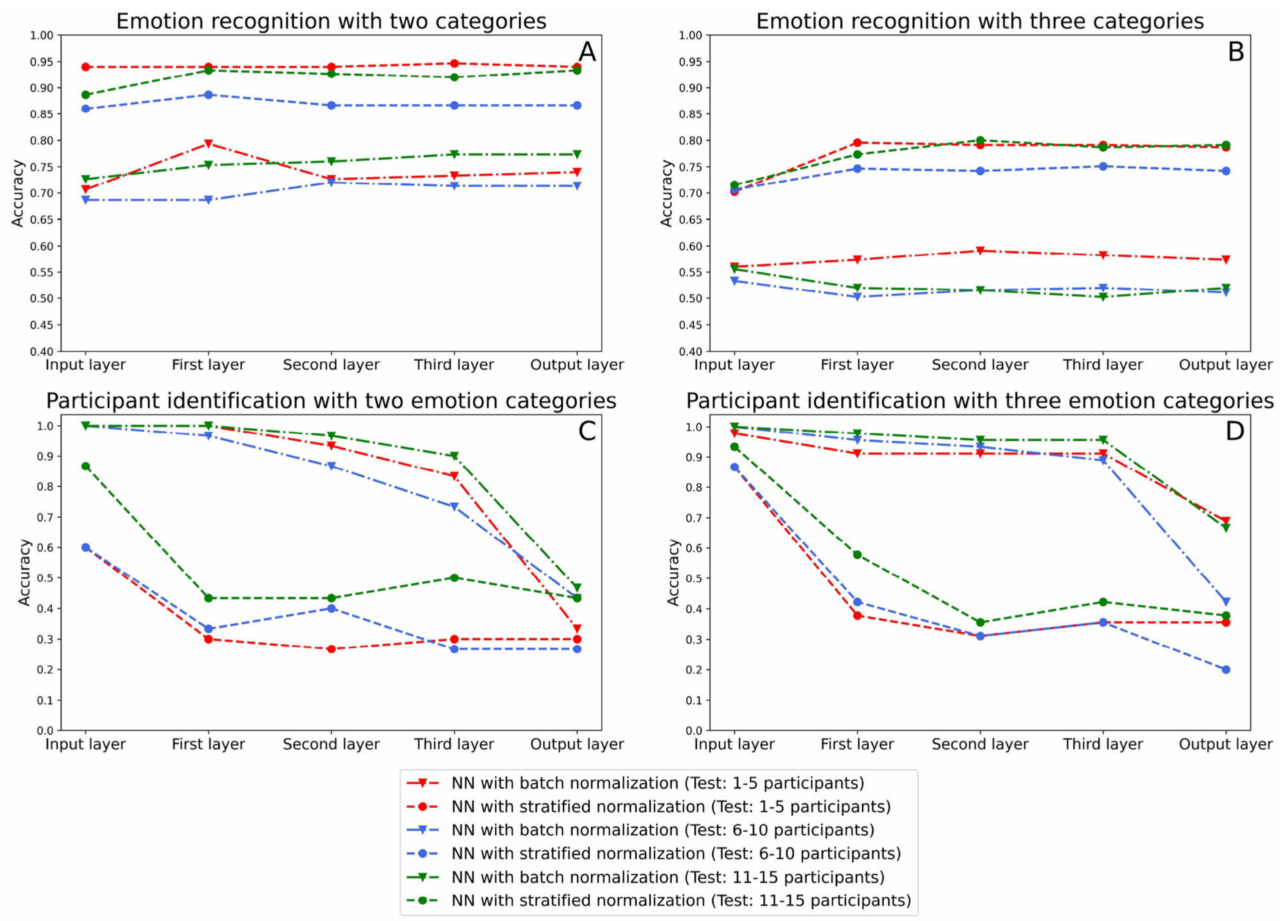

FIGURE 6 | This figure indicates the classification results for the emotion recognition and participant identification in the input, hidden layers, and output of the neural network using multitaper as feature extraction method. **(A,B)** Plot the emotion recognition accuracies, and **(C,D)** show the participant identification accuracy.

each layer of the neural network. All analyses revealed statistically significant effects of the normalization method [$F_{(1,\,8)} = 20.4$, $p = 0.002$, $\eta_p^2 = 0.72$ for the input layer, $F_{(1,\,8)} = 217.6$, $p < 0.001$, $\eta_p^2 = 0.97$ for the first layer, $F_{(1,\,8)} = 352.1$, $p < 0.001$, $\eta_p^2 = 0.98$ for the second layer, $F_{(1,\,8)} = 118.9$, $p < 0.001$, $\eta_p^2 = 0.94$ for the third layer, $F_{(1,\,8)} = 8.83$, $p = 0.018$, $\eta_p^2 = 0.53$ for the output layer]. This confirms that the normalization method shows an effect for every layer's depths, including for the input and the output layers.

Hence, (1) both normalization methods significantly reduce the participant identification information, (2) the layers where the participant identification information is significantly reduced varies depending on the normalization method (output layer for batch normalization and first layer for stratified normalization), and (3) the stratified normalization overpasses batch normalization, having an accuracy of $M = 0.33$, SD $= 0.072$ for two categories and $M = 0.31$, SD $= 0.079$ for three categories in the last layer, where the chance level is 0.2 since we are classifying five participants.

To conclude, as hypothesized, the decrease of participant identification accuracy observed for both normalization methods confirms that the brain signature is effectively suppressed throughout the Neural Network, preventing the SVM classifiers from recognizing which participant their input data belongs to. Still, we can still see that some participant identification information remains in the output of the models. Indeed, if there wasn't, then the accuracy of the participant identification would be at a chance level of 20% (considering 5 participants), but instead, it is still at 33% ($M = 0.33$, SD $= 0.072$) for two categories and 31% ($M = 0.31$, SD $= 0.079$) for three categories in the last layer of the models with stratified normalization. This difference is much higher with batch normalization, 41% ($M = 0.41$, SD $= 0.057$) for two categories and 59% ($M = 0.59$, SD $= 0.012$) for three categories.

We are now wondering about the type of mechanism that operates the brain signature suppression in the model with stratified normalization, and specifically about how it affects the data.

3.3.3. Visualization on the Input and Output Layers of the Neural Network

To have a clearer visualization of the classifier's performance, we run the dimensionality reduction tool UMAP (McInnes et al., 2018) to reduce to two dimensions all the predicted values for the input and output layers of the neural network.

Figures 7, **8** show the embedding of the predicted values at the input layer of the neural network. We have established in the section above that a significant amount of brain signature is already suppressed by this layer. Interestingly, the UMAP for batch normalization shows a handful of compact clusters. These clusters match relatively well the participant numbers, but not the emotional rating. Conversely, the UMAP for the stratified normalization is much more spread out along both dimensions, and neither the emotion ratings nor the participant numbers are recognizable in the cloud of embeddings. This difference suggests that the stratified normalization method induces a redistribution of the activations in output of the first layer in a more spread space of representation, possibly to facilitate further processing in the rest of the network. Then the spread would cause the embeddings to lose their information about the participant number.

Figures 9, **10** show the embedding of the predicted values at the output layer of the neural network. Our previous results showed that at the output layer, the emotion recognition accuracy is higher and the participant identification accuracy is lower for models trained with stratified normalization. Indeed, this time, embeddings are more compact on the UMAP for the stratified normalization rather than for the batch normalization, and easily recognizable for the emotion ratings rather than for the participant numbers—for which the spread of colors seems to indicate that most of the brain signature is gone indeed.

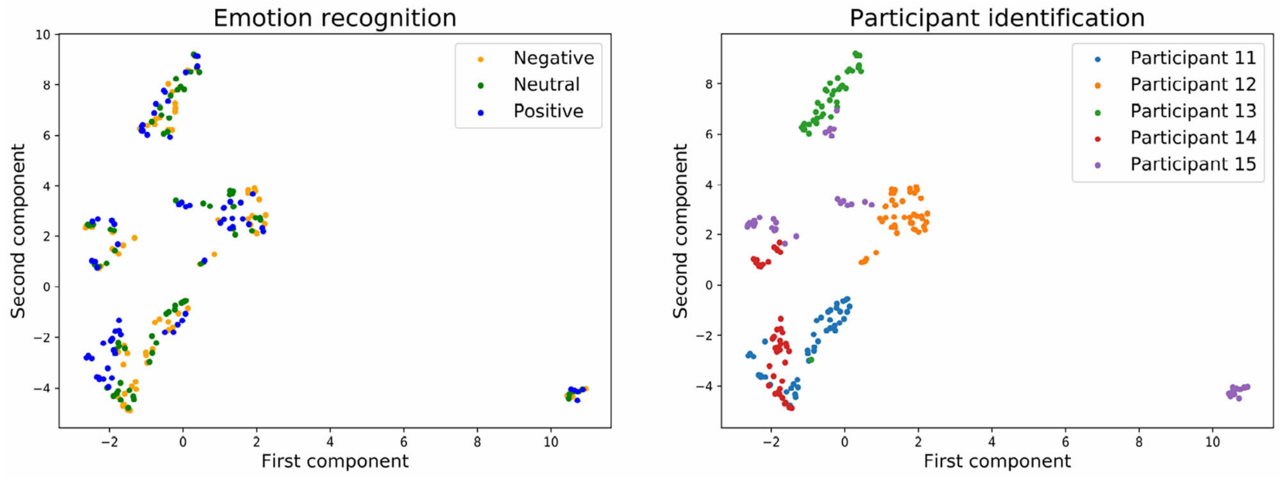

FIGURE 7 | Embedding of the input of the neural network with batch normalization and with three emotion categories.

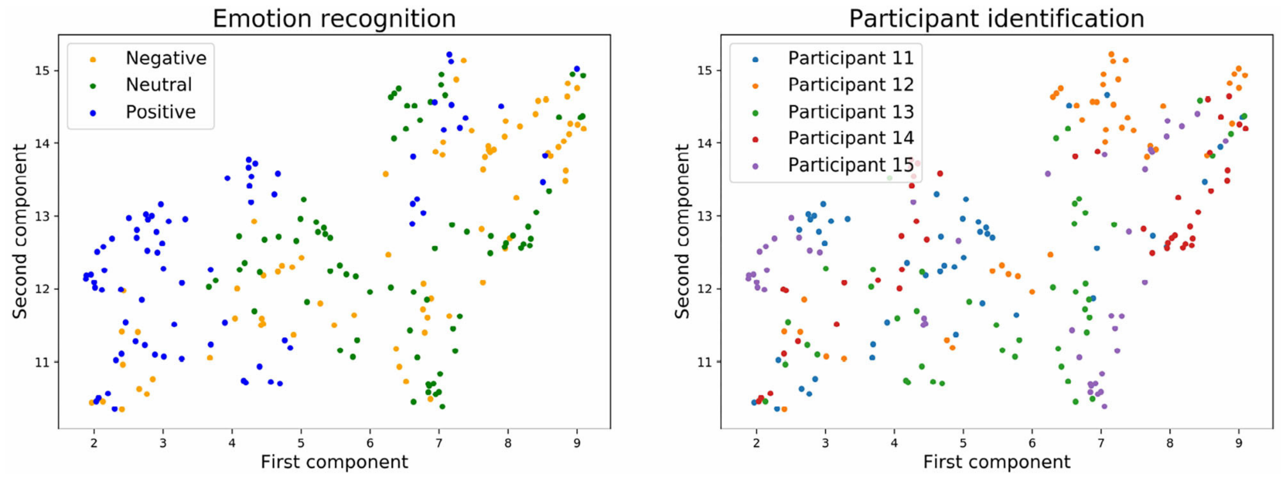

FIGURE 8 | Embedding of the input of the neural network with stratified normalization and with three emotion categories.

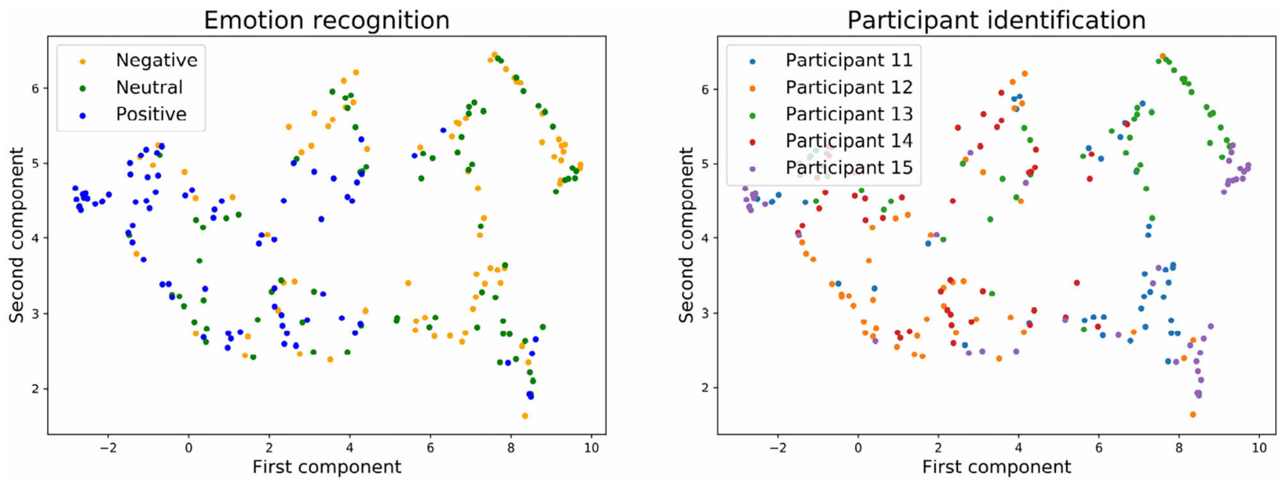

FIGURE 9 | Embedding of the output of the neural network with batch normalization and with three emotion categories.

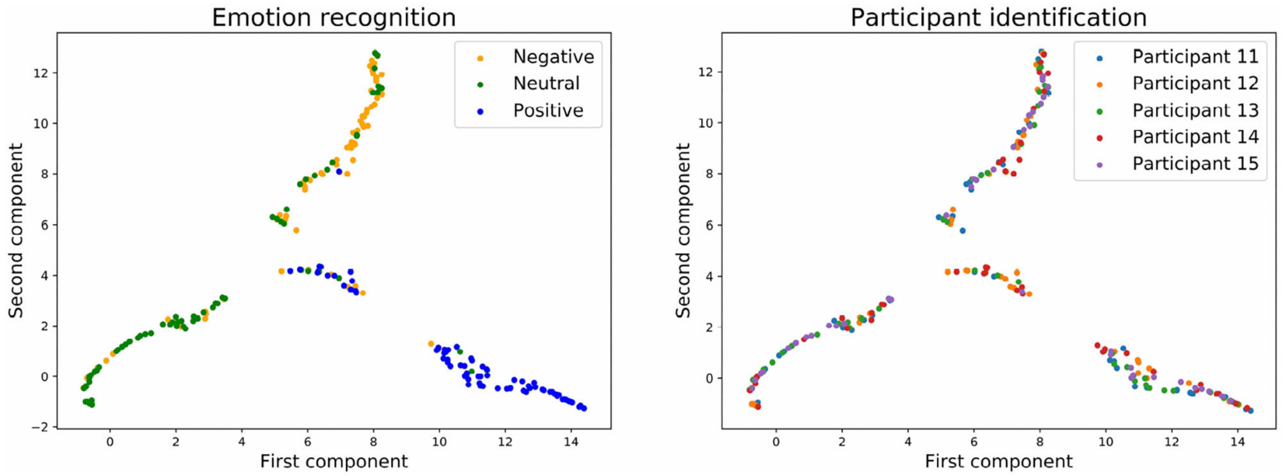

FIGURE 10 | Embedding of the output of the neural network with stratified normalization and with three emotion categories.

4. CONCLUSION

In recent years, researchers have introduced and evaluated different approaches that permit to build robust participant-independent models without the need for prior recorded data from each participant. Specifically, the primary focus is on finding features that do not vary across participants. However, since these methods still present lower accuracy than participant-dependent models, researchers are also investigating other approaches, such as data normalization. In this study, we propose and evaluate a new participant-based feature normalization method, named *stratified normalization*, to improve the cross-subject emotion recognition accuracy of participant-independent models.

The evaluation of this method has been carried out by setting an experiment where we recorded the effects of three independent variables (labeling method, normalization method, and feature extraction method) onto the cross-subject emotion recognition accuracy. The selected dataset for this analysis has been the SEED dataset, where the brainwaves of 15 participants were recorded while watching the same 15 film clips across three different sessions.

We first compared the Welch, multitaper, and differential entropy methods for extracting features in task of binary and ternary classification. Our participant-independent model was a CNN-based network with an input and three hidden layers each followed by our new, stratified normalization method, and an output followed by a softmax function for classification. The highest leave-one-out cross-validation mean accuracy with our model was $M = 0.916$, SD = 0.074 for binary classification and $M = 0.796$, SD = 0.104 for ternary classification, when extracting the features with the multitaper method. We also compared our stratified normalization method with batch normalization, obtaining after implementing an ANOVA test that the classification accuracies for stratified normalization was statistically higher than batch normalization. We also observed that including the neutral labels in the model hinders the classification of the negative labels, decreasing their classification accuracy from 90.67 to 75.11%.

Then, we found out that implementing stratified normalization is highly efficient in reducing the inter-participant variability from the data. Indeed, by training SVMs to try and recognize which participant the activation data of a given layer belongs to, we could observe that the participant identification information, or brain signature, was lost from a layer to another.

As we compared the embeddings at the level of the input and output layers, we could see that the stratified normalization already erases this brain signature in the input layer, such that by the end of the network, it is almost gone already—33% for two categories and 31% for three categories in the last layer of the models with stratified normalization, approaching a chance level of 20%. It would be interesting to look for new ways of improving this result further.

Regarding the published articles, our method outperforms the rest of the proposed methods for binary classification and overpasses the works that did not use the data for validation during the training process for ternary classification.

These results indicate the high applicability of stratified normalization for cross-subject emotion recognition tasks, suggesting that this method could be applied not only to other EEG classification datasets but also to other applications that require domain adaptation algorithms.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Public dataset available at: http://bcmi.sjtu.edu.cn/home/seed/seed.html. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JF ran the experiment, analyzed the results, and wrote the manuscript. NG proposed the idea and was in charge of the technical supervision. OW and AP were the supervisors of the project and provided revision suggestions. All authors contributed to the article and approved the submitted version.

REFERENCES

Arevalillo-Herráez, M., Cobos, M., Roger, S., and García-Pineda, M. (2019). Combining inter-subject modeling with a subject-based data transformation to improve affect recognition from EEG signals. *Sensors* 19. doi: 10.3390/s19132999

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv arXiv:1607.06450*.

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K. M., and Robbins, K. A. (2015). The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Front. Neuroinform*. 9:16. doi: 10.3389/fninf.2015.00016

Brave, S., and Nass, C. (2009). "Emotion in human-computer interaction," In: A. Sears and J. Jacko (Eds.), *Human-Computer Interaction Fundamentals* (p. 53–68). doi: 10.1201/b10368-6

Cao, R., Hao, Y., Wang, X., Gao, Y., Shi, H., Huo, S., et al. (2020). EEG functional connectivity underlying emotional valance and arousal using minimum spanning trees. *Front. Neurosci*. 14:355. doi: 10.3389/fnins.2020.00355

Chai, X., Wang, Q., Zhao, Y., Li, Y., Liu, D., Liu, X., et al. (2017). A fast, efficient domain adaptation technique for cross-domain electroencephalography(EEG)-based emotion recognition. *Sensors* 17, 1–21. doi: 10.3390/s17051014

Chai, X., Wang, Q., Zhao, Y., Liu, X., Bai, O., and Li, Y. (2016). Unsupervised domain adaptation techniques based on auto-encoder for non-stationary EEG-based emotion recognition. *Comput. Biol. Med*. 79, 205–214. doi: 10.1016/j.compbiomed.2016.10.019

Chang, Y. W., Hsieh, C. J., Chang, K. W., Ringgaard, M., and Lin, C. J. (2010). Training and testing low-degree polynomial data mappings via linear SVM. *J. Mach. Learn. Res*. 11, 1471–1490. doi: 10.5555/1756006.1859899

Chen, D. W., Miao, R., Yang, W. Q., Liang, Y., Chen, H. H., Huang, L., et al. (2019). A feature extraction method based on differential entropy and linear discriminant analysis for emotion recognition. *Sensors* 19:1631. doi: 10.3390/s19071631

Cimtay, Y., and Ekmekcioglu, E. (2020). Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset EEG emotion recognition. *Sensors* 20:2034. doi: 10.3390/s20072034

Craik, A., He, Y., and Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: a review. *J. Neural Eng*. 16:31001. doi: 10.1088/1741-2552/ab0ab5

De Andrés, I., Garzón, M., and Reinoso-Suárez, F. (2011). Functional anatomy of non-REM sleep. *Front. Neurol*. 2:70. doi: 10.3389/fneur.2011.00070

Duan, R. N., Zhu, J. Y., and Lu, B. L. (2013). "Differential entropy feature for EEG-based emotion classification," in *International IEEE/EMBS Conference on Neural Engineering* (Boston, MA). doi: 10.1109/NER.2013.6695876

Dzedzickis, A., Kaklauskas, A., and Bucinskas, V. (2020). Human emotion recognition: review of sensors and methods. *Sensors* 20, 1–41. doi: 10.3390/s20030592

El Keshky, M. E. S. (2018). Emotion dysregulation in mood disorders: a review of current challenges. *J. Psychol. Clin. Psychiatry* 9:585. doi: 10.15406/jpcpy.2018.09.00585

Fels, S., El-Nasr, M. S., Graham, N., Anacleto, J., Kapralos, B., and Stanley, K. (2011). *Preface*. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5–6.

Gery, I., Miljkovitch, R., Berthoz, S., and Soussignan, R. (2009). Empathy and recognition of facial expressions of emotion in sex offenders, non-sex offenders and normal controls. *Psychiatry Res*. 165, 252–262. doi: 10.1016/j.psychres.2007.11.006

Heuer, K., Rinck, M., and Becker, E. S. (2007). Avoidance of emotional facial expressions in social anxiety: the approach-avoidance task. *Behav. Res. Ther*. 45, 2990–3001. doi: 10.1016/j.brat.2007.08.010

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv arXiv:1502.03167*.

Issa, M. F., and Shedeed, H. A. (2016). Brain-EEG signal classification based on data normalization for controlling a robotic arm. *Int. J. Tomogr. Simul*. 29, 72–85. Available online at: http://www.ceser.in/ceserp/index.php/ijts/article/view/3990

Jatupaiboon, N., Pan-Ngum, S., and Israsena, P. (2013). Real-time EEG-based happiness detection system. *Sci. World J.* 2013:618649. doi: 10.1155/2013/618649

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv:1412.6980*.

Koelstra, S., Mühl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., et al. (2012). DEAP: A database for emotion analysis; Using physiological signals. *IEEE Trans. Affect. Comput*. 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Lan, Z., Sourina, O., Wang, L., Scherer, R., and Muller-Putz, G. R. (2019). Domain adaptation techniques for EEG-based emotion recognition: a comparative study on two public datasets. *IEEE Trans. Cogn. Dev. Syst*. 11, 85–94. doi: 10.1109/TCDS.2018.2826840

Li, X., Song, D., Zhang, P., Zhang, Y., Hou, Y., and Hu, B. (2018). Exploring EEG features in cross-subject emotion recognition. *Front. Neurosci*. 12:162. doi: 10.3389/fnins.2018.00162

Li, X., Zhao, Z., Song, D., Zhang, Y., Pan, J., Wu, L., et al. (2020). Latent factor decoding of multi-channel EEG for emotion recognition through autoencoder-like neural networks. *Front. Neurosci*. 14:87. doi: 10.3389/fnins.2020.00087

Logesparan, L., Casson, A. J., and Rodriguez-Villegas, E. (2011). "Assessing the impact of signal normalization: preliminary results on epileptic seizure detection," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Boston, MA), 1439–1442. doi: 10.1109/IEMBS.2011.6090356

Mansouri, A., and Castillo-Guerra, E. (2019). Multitaper MFCC and normalized multitaper phase-based features for speaker verification. *SN Appl. Sci*. 1:290. doi: 10.1007/s42452-019-0305-y

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*. doi: 10.21105/joss.00861

Milligan, G., and Cooper, M. (1988). A study of standardization of variables in cluster analysis. *J. Classif*. 5, 181–204. doi: 10.1007/BF01897163

Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., et al. (2018). A review of emotion recognition using physiological signals. *Sensors* 18:2074. doi: 10.3390/s18072074

Soleymani, M., Lichtenauer, J., Pun, T., and Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affect. Comput*. 3, 42–55. doi: 10.1109/T-AFFC.2011.25

Song, T., Zheng, W., and Song, P., C. Z. (2020). EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affect. Comput*. 11, 532–541. doi: 10.1109/TAFFC.2018.2817622

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2017). Instance normalization: the missing ingredient for fast stylization. *arXiv:1607.08022*.

Welch, P. D. (1975). The use of fast Fourier transform for the estimation of power spectra. *Digit. Signal Process*. 532–574.

Yang, F., Zhao, X., Jiang, W., Gao, P., and Liu, G. (2019). Multi-method fusion of cross-subject emotion recognition based on high-dimensional EEG features. *Front. Comput. Neurosci*. 13:53. doi: 10.3389/fncom.2019.00053

Yang, H. Y., Han, J. H., and Min, K. (2019). Distinguishing emotional responses to photographs and artwork using a deep learning-based approach. *Sensors* 19:5533. doi: 10.3390/s19245533

Yin, Z., Wang, Y., Liu, L., Zhang, W., and Zhang, J. (2017). Cross-subject EEG feature selection for emotion recognition using transfer recursive feature elimination. *Front. Neurorobot*. 11:19. doi: 10.3389/fnbot.2017.00019

Yousif, E. S., Abdulbaqi, A. S., Hameed, A. Z., and Al-din M. N. S. (2020). Electroencephalogram signals classification based on feature normalization. *IOP Conf. Ser*. 928:032028. doi: 10.1088/1757-899X/928/3/032028

Zhang, W., Wang, F., Jiang, Y., Xu, Z., Wu, S., and Zhang, Y. (2019). *Cross-Subject EEG-Based Emotion Recognition With Deep Domain Confusion*. Cham: Springer. doi: 10.1007/978-3-030-27526-6_49

Zheng, W. L., Liu, W., Lu, Y., Lu, B. L., and Cichocki, A. (2019). EmotionMeter: a multimodal framework for recognizing human emotions. *IEEE Trans. Cybernet*. 49, 1110–1122. doi: 10.1109/TCYB.2018.2797176

Zheng, W. L. and Lu, B. L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Mental Dev*. 7, 162–175. doi: 10.1109/TAMD.2015.2431497

Zhu, D., Yao, H., Jiang, B., and Yu, P. (2020). "Negative log likelihood ratio loss for deep neural network classification," in *Proceedings of the Future Technologies Conference (FTC) 2019* (Cham: Springer).

Conflict of Interest: JF, NG, OW, and AP were employed by company Cross Compass Ltd.