Explainable Classification of EEG Data for an Active Touch Task using Shapley Values

Haneen Alsuradi[1][0000−0001−7396−444X], Wanjoo Park[2][0000−0003−1467−4156], and Mohamad Eid[2][0000−0002−6940−7891]

[1] New York University, Tandon School of Engineering, NY 11201, USA
[2] Engineering Division, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, 129188
{haneen,wanjoo,mohamad.eid}@nyu.edu

Abstract. Machine learning has been used in the last decade to solve many problems in the haptics field. In particular, EEG data that is recorded during haptic interactions was used to train machine learning (ML) models to answer questions that are of interest to the neurohaptics community. However, the behavior of machine learning models in taking out their decisions is treated as black box hindering the interpretability of these decisions. In this paper, we used Shapley values, a concept from game theory, to explain the behavior of a tree-based classifier model in classifying electroencephalography data that was collected during an interaction with a surface haptic device under two conditions: with and without tactile feedback. We trained a tree-based ML model to classify data based on the presence or absence of tactile feedback. Using Shapley values, we identified the features (across and within channels) that contribute the most to the classification decision. Results showed channel AF3 and neural activity after 700 ms from the onset contributed the most in recognizing tactile feedback in the interaction. This study demonstrates the use of explainable machine learning in the field of Neurohaptics.

Keywords: Neurohaptics · haptics · Explainable machine learning · EEG.

1 Introduction

There is a growing interest within the haptics community to involve human brain assessment techniques as new tools to understand the human haptic experience. Methods such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) are well-established and can record signals from the brain during haptic interactions [15]. Conventionally, self-reporting is used to assess the human haptic experience. However, brain assessment methods have many advantages that can complement self-reporting in many ways. Participants going under self-reporting are prone to difficulty in expressing themselves [14]. In addition, reporting is usually done once the experiment is over which means

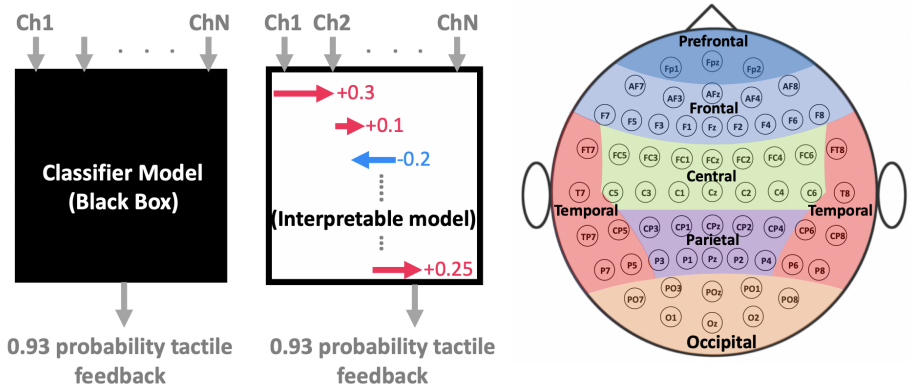

Fig. 1. Left: An interpretable model is capable of giving each of the features a credit in the final prediction probability. Right: Spatial map of EEG channels.

the response is not recorded during the time of interaction. As human memory is susceptible to forgetting or distortion, late reporting can be inaccurate [11]. Brain assessment methods emerged as quantitative and objective complementary measures to overcome the mentioned limitations of self-reporting. fMRI has a high spatial resolution; however, its temporal resolution is in the order of a few seconds which is much slower than a typical neural process [10], [6]. Moreover, fMRI imposes technical challenges in incorporating electronics within its vicinity due to the extremely high magnetic field. On the other hand, EEG is a more affordable apparatus that measures the brain's electrical activity with a high temporal resolution, making it particularly suitable for understanding the temporal aspect of the neural processes. Additionally, electronic devices which might be part of the haptic interaction are easily accommodated in EEG-based experiments.

In the past few years, EEG data was not only used to unveil information about the neural processes during a haptic interaction, but also was used to train machine learning (ML) models to answer questions that are of interest to the neurohaptics community. For example, ML models were employed to classify objects with different physical and geometrical properties through grasping tasks or tactile exploration using EEG data [3], [9]. ML models were also used in affective haptics field, for example to recognize affective haptic stimuli conveyed by different fabrics or determine the degree of pleasure level during an interpersonal interaction using EEG data [7], [16].

However, the behavior of machine learning models in taking out their decisions is treated as a black box hindering the interpretability of these decisions. Understanding why an ML model makes a certain prediction can be as important as the prediction itself. Thus, it is of importance to involve explainable machine learning (XML) to serve the neurohaptics and the HCI community that uses ML to classify brain activation recorded during a human-computer interaction. XML makes ML models more transparent by justifying the classifier predictions

and accrediting each of the features with an importance score in making the prediction, thus improving our understanding of the psychophysics of the task.

Our previous work showcased the use of a support vector machine (SVM) classifier in detecting the presence of tactile feedback during interaction with a touchscreen device through EEG data [1]. In this paper, we explain a tree-based classifier model in predicting the presence or absence of tactile feedback using Shapley values [17]. Figure 1 (left) illustrates the idea behind this work; an explanation is given by crediting each of the channels/features with a contribution score in predicting the probability of the presence or absence of the tactile feedback.

2 Experimental study

2.1 Experiment design

In this study, we utilize the EEG data from our previous work [15]. The experiment consisted of an active touch task in which participants were asked to slide their index finger across guitar strings displayed on a Tanvas touchscreen device from predefined start to end locations. The screen is capable of providing friction-based tactile feedback which is turned on and off thus having two types of stimulation modes; one mode is activated per trial. The order of stimulation was randomized while considering the "counterbalancing" paradigm. That is to say, participants are divided in half such that one half performs the two conditions in one order and the other half performs the conditions in the reverse order. Visual (shaken strings) and auditory feedback (guitar sound), however, were always provided. Neural activation was recorded during the haptic interaction using a 64-channel EEG system; electrode locations are shown in Figure 1 right. Participants were trained such that the interaction time with the touchscreen device in one trial would take around 1000ms. A number of 96 trials per condition (with or without tactile feedback) were conducted for each participant. Twenty-six participants were recruited for this study. The study was carried out with an approved protocol by New York University Abu Dhabi Institutional Review Board (IRB: #073-2017).

2.2 EEG data processing

EEG signals were first down-sampled from 2500 Hz to 1250 Hz and band pass filtered (0.1–55 Hz). After discarding eye-movement and muscle artifacts using the artifact subspace reconstruction, EEG data was epoched and divided into two categories depending on the presence or absence of the tactile feedback. Power spectral densities (PSD) of the frequency bands (theta, alpha, beta and gamma) were then calculated. A thorough analysis of the differences in PSD between the two stimulation modes was carried out in our previous work [15]; it was found that beta band power was significantly higher during the presence of tactile feedback on multiple locations including the ipsilateral-parietal, contralateral-parietal, middle-parietal and middle-frontal regions.

3 Proposed Method

Two ML classifiers were created; classifier 1 takes its inputs from the 64 EEG channels in order to predict the presence or absence of the tactile feedback. Shapley values were used to evaluate the most influential channel in the prediction process. Once identified, classifier 2 is trained using the EEG data solely from the identified channel. Shapley values are again employed to identify the timestamp/time-period during which the neural activation of the identified channel is most influential in the prediction process. For both models, we trained an XGBoost model which is a tree-based classifier. Below, the proposed method will be explained in detail.

3.1 Feature Extraction

For classifier 1, since the data is high-dimensional (from 64 channels), a data reduction/feature selection method is needed. After extracting the beta band PSD for every channel, we examined the grand average plots (mean over trials and subjects) of the channels across the scalp . Figure 2 shows an example grand average PSDs from channels F1 and POz from the middle frontal cortex and middle parietal cortex, respectively. It can be noticed that the peak amplitude and latency combined can be a representative feature for each channel. The use of peak as a feature is commonly used in PSD analysis [8] as well as ERP analysis [5]. We thus defined a feature for each channel by multiplying its peak amplitude value with its corresponding latency; we call it *peak-factor*. Due to the small number of observations (i.e. participants) we used a time-shifting data

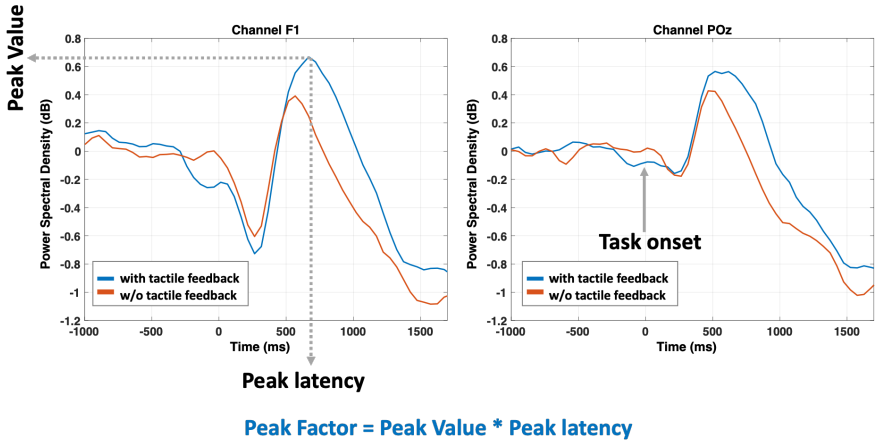

Fig. 2. Grand average PSD plots from two channels in beta band, F1 (frontal area) and POz (occipital area) showing the distinctive features between the two stimulation modes.

augmentation scheme in order to populate the training data, and hence improve the accuracy of the model. In the time-shifting scheme, each PSD signal (i.e.: each trial) was shifted in time 50 ms forward and backward. Thus, the size of the training data has tripled in size and hence the accuracy of the classifier has improved.

3.2 Classifier

As mentioned earlier, two ML classifiers were trained. Classifier 1 was trained on the features extracted above, namely the *peak-factors* from the 64 channels. Once Shapley values are extracted for each of the 64 channels, the highest Shapley value corresponds to the most influential channel in the prediction decision. Classifier 2 on the other hand was trained feeding the full waveform of the beta-band PSDs of the most influential channel (instead of the feeding the *peak-factor* of the waveform) of all subjects under the two stimulation modes.Shapley values identified the timestamp/duration at which the neural activity of the channel was most influential in differentiating between the two haptic modes. For both classifiers, an extreme gradient boosting (XGBoost) model was trained to predict the class of the stimulation mode. XGBoost is an optimized decision-tree ensemble ML algorithm that has been widely accepted and recognized in the last few years [2]. A single decision-tree model suffers from high variance, which means the model tends to overfit to the training data [4]. An ensemble-of-trees model on the other hand, such as random forest, is based on growing trees randomly to reduce the variance. For further optimization, additive training (boosting) method can be used in growing trees such that each tree tries to resolve the deficiencies of the previous tree. XGBoost classifier implements the boosting technique with improved performance and accuracy [2]. Data was randomly split into 80% training and 20% testing and XGBoost model was trained and tested accordingly with 84% prediction accuracy for both of the classifiers.

3.3 Shapley values

Shapley values, a concept from game theory, is a credit attribution method for a player in a game. Shapley values were first used in machine learning as part of a unified framework (named SHAP) for interpreting predictions such that the game is replaced by the model and the player is replaced by the features of the model. SHAP is used in ML to explain the contribution of each feature in the prediction of the model [13]. SHAP has been successfully used in ML models in the medical domain [12]. In this work, we use SHAP as a tool to explain a tree-based model in the neurohaptics field. SHAP is a local feature attribution method which means that SHAP accredits each feature with a contribution score given a single sample/trial input data. In other words, SHAP is designed to explain a prediction $f(x)$ based on a single input vector x. A global insight, however, can be extracted once all the local explanations are found. To calculate the Shapley value of a feature i, sets of all the possible combinations of the n features are created, excluding the ith feature. The model f is evaluated with

$(f(S \bigcup \{i\}))$ and without $f(S)$ feature i and the difference in the prediction for the input data x is calculated. The input data *(*x) will contain *(S+1)* features when fed to $(f(S \bigcup \{i\}))$ model while *(S)* features in the $f(S)$ model. The difference in the prediction is the marginal contribution of the feature i in prediction. This process is repeated for all the other formed features combinations. Thus, Shapley value for a specific feature i is calculated by finding the average of the marginal contribution across all possible permutation of features' combinations. The equation below is used to calculate the Shapley value for feature i under the model function f:

$$\varphi_{i(f)} = \sum_{S \subseteq N/\{i\}} \frac{|S|!(n-|S|-1)!}{n!} (f_{S \bigcup \{i\}}(x_{S \bigcup i}) - f_S(x_S)) \tag{1}$$

where N is the set of all features, S is a subset of the features without feature i, n is the number of all features, $|S|$ is the cardinality of S (which is simply the number of elements in S for a finite set) and f is the function of the classifier model. Note that the second term in the bracket calculates the marginal contribution of feature i by considering a feature set with and without feature i and subtracting their predictions. The term before the bracket represents the number of all the possible ways of forming combinations per S divided by the total number of possible permutations.

4 Results and Discussion

As mentioned before, Shapley values are effective in revealing the impact of an input feature on an individual prediction; this is called local explanation. Combining many local explanations can lead to a global insight into the model's behavior [12].

4.1 Channel level explanation

To understand which of the EEG channels are the most influential for the XGBoost classifier 1 during prediction, we plot the Shapley values of each feature (channel's *peak-factor*) for all the trials in a beeswarm plot as shown in Figure 3 (right). The channels are ordered with respect to importance (i.e. AF3 is the most important). Each dot represents a specific channel's *peak-factor* for a single trial (instance) in the training data. The color of the dot corresponds to the *peak-factor* value. High *peak-factor* values are colored in red while lower *peak-factor* values are colored in blue. The horizontal location of the dot on the other hand corresponds to the Shapley value of the feature; it explains whether the effect of that feature is associated with a positive or negative contribution to the prediction probability of the presence or absence of tactile feedback. From this plot, it can be observed that lower *peak-factor* values in AF3 channel, for example, will contribute negatively to the prediction probability of having tactile feedback. Another observation is that sometimes, a low AF3 *peak-factor* can greatly

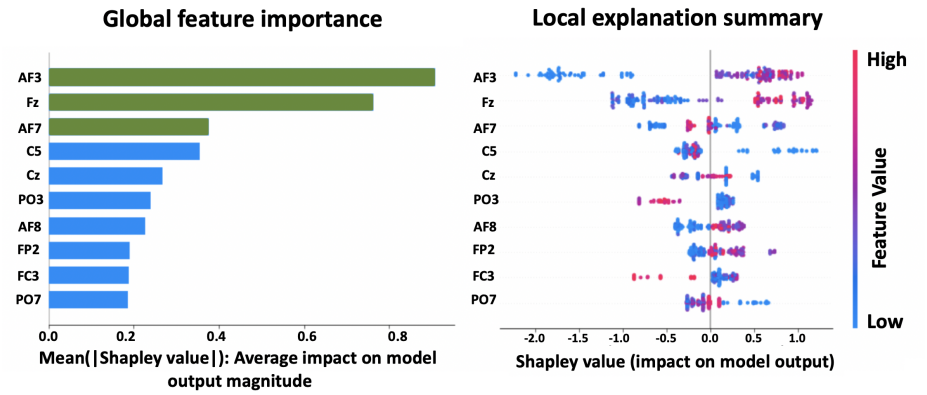

Fig. 3. Left: Global feature importance indicates that the channels in the middle frontal cortex, namely (AF3, Fz and AF7) contribute the most in the model's prediction. Right: Beeswarm plot showing the impact of each channel's *peak-factor* on the prediction probability. Each dot represents a sample.

reduce the prediction probability of having tactile feedback, much more than a high *peak-factor* would increase the prediction probability of having tactile feedback. The global feature importance on the other hand is shown in Figure 3 (left). For each channel, the mean absolute value of the Shapley values is plotted. Middle frontal electrodes rank the top in implying the presence/absence of tactile feedback. Another way to explore a single trial prediction explanation is

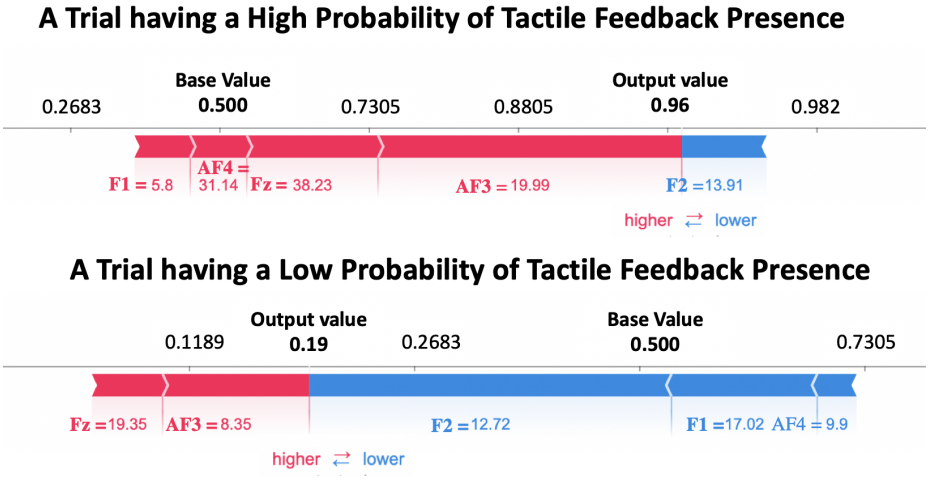

Fig. 4. Two different trials showing how the *peak-factor* of the shown channels contribute positively (red) and negatively (blue) in the prediction process.

through what is called the force plot shown in Figure 4. The figure shows two different examples of trials, one with high probability of the presence of tactile feedback and one with low probability of the presence of tactile feedback. In each force plot, each of the channels contribution in pushing the prediction probability from the base value (0.5) is illustrated in magnitude and direction (i.e.: an indication of channels' correlation to presence/absence of tactile feedback). Note that only 5 of the channels' contributions are shown in the force plot for illustration purposes. From a neural perspective, beta activity at the middle frontal cortex is associated with an increased cognitive processing [5]. This is a possible indication that tactile feedback results in a more immersive interaction as it resembles reality [15].

4.2 Activity within-channel explanation

Since it is found that AF3 channel is the most influential channel in classifier 1, we would like to obtain further explainability and find the most impactful timestamp at which the neural activity of AF3 is important. As mentioned earlier, classifier 2 was trained using the full waveform of the beta-band PSDs of the most influential channel instead of the feeding the *peak-factor* of the waveform. Each value in the waveform at each timestamp is considered as a feature. The local explanation summary plot shown in Figure 5 (right). The figure shows that neural activity after 700 ms from the onset of the task contribute the most in the prediction probability in a descending order with time. It can also be observed that the feature at 718 ms produces a strong force to pull up or down the prediction probability depending on the feature value (no sample points have a zero Shapley value). Additionally, long tails along the x-axis in the same figure (such as at feature 819 ms), indicate that for some individuals, this feature is extremely important in impacting the prediction probability. Note that due to EEG data digitization, a specific timestamp is not important per se, instead, a group of consecutive timestamps indicate that this period of time (after the

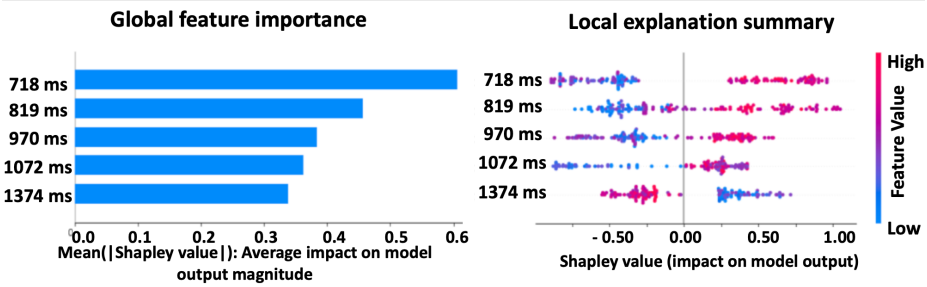

Fig. 5. Left: Global feature importance indicate timestamps after 700 ms from the onset of the task contribute the most in the model's prediction. Right: Beeswarm plot showing the impact of neural activation (at the indicated timestamps) of AF3 channel on the prediction probability. Each dot represents a sample.

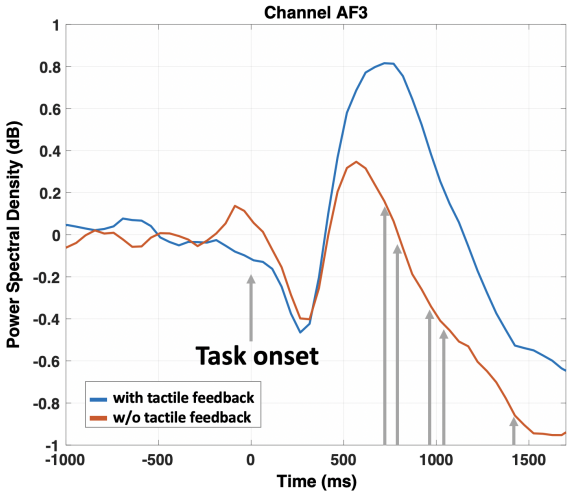

Fig. 6. Grand average PSD plots from AF3 channel with marked important features.

onset of the stimulus) is important in distinguishing the two classes apart. Figure 5 (left) shows the global feature importance of classifier 2. The five most influential timestamps are indicated on the grand average PSD of AF3 under the two stimulation modes, shown in Figure 6.

5 Conclusion

In this paper, we demonstrated the use of a game theoretic concept, Shapley value, in explaining the behavior of a tree-based classifier (XGBoost). The classifier was trained on EEG data to predict the presence or absence of tactile feedback during interaction with a touchscreen device. We found that the channel AF3 located in the middle frontal cortex contributes the most in the decision making of the classifier. We could also demonstrate explanations of a specific sample prediction and the contribution of each channel in making the prediction. We further showed that Shapley values provided an interpretation of the classifier behavior by finding the most influential timestamps at which the neural activity is important towards classification. Neural activity after 700 ms from the onset contributed the most. These results are consistent with those found exploratively in previous studies [15]. Therefore, we believe that EEG channels and time periods that contribute the most in classifications found through Shapley values will assist researchers in exploring meaningful features in experiments in neurohaptics and HCI.

Acknowledgement

This research was funded by NYU Abu Dhabi PhD Fellowship Program.

References

1. Alsuradi, H., Pawar, C., Park, W., Eid, M.: Detection of tactile feedback on touch-screen devices using eeg data. In: 2020 IEEE Haptics Symposium (HAPTICS). IEEE (2020)
2. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
3. Cisotto, G., Guglielmi, A.V., Badia, L., Zanella, A.: Classification of grasping tasks based on eeg-emg coherence. In: 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom). pp. 1–6. IEEE (2018)
4. Dietterich, T.G., Kong, E.B.: Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Tech. rep., Technical report, Department of Computer Science, Oregon State University (1995)
5. Egner, T., Gruzelier, J.H.: Eeg biofeedback of low beta band components: frequency-specific effects on variables of attention and event-related brain potentials. Clinical neurophysiology **115**(1), 131–139 (2004)
6. Glover, G.H.: Overview of functional magnetic resonance imaging. Neurosurgery Clinics **22**(2), 133–139 (2011)
7. Greco, A., Nardelli, M., Bianchi, M., Valenza, G., Scilingo, E.P.: Recognition of affective haptic stimuli conveyed by different fabrics sing eeg-based sparse svm. In: 2017 IEEE 3rd international forum on research and technologies for society and industry (RTSI). pp. 1–5. IEEE (2017)
8. Grummett, T.S., Fitzgibbon, S.P., Lewis, T.W., DeLosAngeles, D., Whitham, E.M., Pope, K.J., Willoughby, J.O.: Constitutive spectral eeg peaks in the gamma range: suppressed by sleep, reduced by mental activity and resistant to sensory stimulation. Frontiers in human neuroscience **8**, 927 (2014)
9. Khasnobish, A., Konar, A., Tibarewala, D., Bhattacharyya, S., Janarthanan, R.: Object shape recognition from eeg signals during tactile and visual exploration. In: International Conference on Pattern Recognition and Machine Intelligence. pp. 459–464. Springer (2013)
10. Kim, S.G., Richter, W., Uğurbil, K.: Limitations of temporal resolution in functional mri. Magnetic resonance in medicine **37**(4), 631–636 (1997)
11. Loftus, E.F., Pickrell, J.E.: The formation of false memories. Psychiatric annals **25**(12), 720–725 (1995)
12. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From local explanations to global understanding with explainable ai for trees. Nature machine intelligence **2**(1), 2522–5839 (2020)
13. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in neural information processing systems. pp. 4765–4774 (2017)
14. Morin, C.: Neuromarketing: the new science of consumer behavior. Society **48**(2), 131–135 (2011)
15. Park, W., Jamil, M.H., Eid, M.: Neural activations associated with friction stimulation on touch-screen devices. Frontiers in neurorobotics **13**, 27 (2019)
16. Saha, A., Konar, A., Bhattacharya, B.S., Nagar, A.K.: Eeg classification to determine the degree of pleasure levels in touch-perception of human subjects. In: 2015 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2015)
17. Shapley, L.S.: A value for n-person games. Contributions to the Theory of Games **2**(28), 307–317 (1953)