Applying Conformal Prediction to control an exoskeleton

Charalambos Eliades PAMBOSELIADES@HOTMAIL.COM
Harris Papadopoulos H.PAPADOPOULOS@FREDERICK.AC.CY
Computational Intelligence (COIN) Research Lab

Editor: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo and Evgueni Smirnov

Abstract

This paper investigates the use of the Conformal Prediction (CP) framework for providing confidence measures to assist a Brain Machine Interface (BMI) in the task of controlling an exoskeleton using electroencephalogram (EEG) and electrooculogram (EOG) clips. Reliable and accurate control of assistive robotics is still an important challenge because of the noisy nature of EEG's and EOG's and the fact that any misclassification can lead to unwanted actions and serious safety risks. Therefore a technique that will compliment predictions with a well-calibrated indication of how correct they are, should be very beneficial for the particular application as it can significantly enhance safety. Our approach consists of an Inductive Conformal Predictor (ICP) built on top of a Bidirectional Long Short Term Memory (BiLSTM) Neural Network. We conduct experiments on a dataset consisting of EEG and EOG data collected from one subject with a high spinal cord lesion.

Keywords: EEG, EOG, Confidence, Credibility, Brain Machine Interface(BMI), Safety

1. Introduction

Many diseases such as stroke, spinal cord injuries, multiple sclerosis and weaknesses of the skeletal muscles have a negative impact on the daily activities and quality of life of the affected individuals. Hence, there is a demand to develop new therapeutic techniques and assistive methods that will enable patients to boost their day to day activity performance and recover the lost or impaired motion control. In addition, a successful technique or device (exoskeleton) will release the therapists from the intensive labor of rehabilitation training (Brown-Triolo et al., 2002).

By the term exoskeleton we mean a mechanical wearable device that is designed to mimic body parts such as ankle joints; when the device is worn, the torque produced by the actuators is transferred to the body (He and Kiguchi, 2007). Exoskeleton robotics can efficiently incorporate the cognitive ability of a human being and can assist users in performing heavy activities. Such devices have been developed into full-limb exoskeletons, that is, upper and lower limb exoskeletons and other exoskeleton robots, to support shoulders, elbows, wrists, and ankle joints. This kind of devices can be controlled by physiological signals, like Electromyography (EMG), or brain signals, i.e. electro-encephalography (EEG). To decode these signals into decisions brain computer interfaces (BCI) or brain-machine interfaces (BMI) are used (Al-Quraishi et al., 2018).

BMI translates EEG into control signals enhancing human machine interaction, e.g. allowing impaired individuals to operate assistive systems such as hand prostheses or exoskeletons. The challenge in these systems is to correctly classify the specific brain signals,

as a misclassification may lead to unwanted actions and dangerous safety risks (Witkowski et al., 2014). Consequently a technique which can provide a reliable indication of how likely individual predictions are of being correct is very desirable. The use of such a technique can enable control of the frequency of unwanted actions, by performing only the actions that are certain at a predefined level of confidence while postponing all uncertain actions until enough evidence for the desired action is observed i.e. additional signals for the same action.

Here we propose a CP which can address the problem described above by controlling the error rate of the underlying technique. In particular CP is able to produce prediction sets that are guaranteed to contain the true class with respect to a confidence level. Therefore in practice if for the desirable confidence level the prediction set contains more than one labels the system can wait for more evidence before performing any action, thus avoiding unwanted actions and ensuring safety.

Our approach consists of the following steps. Discrete Fourier Transformation is applied in each EEG. Then we feed a Bidirectional Long Short Term Memory (BiLSTM) Neural Network with the amplitudes of the EEG clip or the amplitudes of the EEG clip together with an EOG clip which will provide us with class probabilities. Finally we apply CP on top of these probabilities to produce prediction sets and confidence measures.

The rest of the paper starts with an overview of related work on BMI for controlling exoskeletons. Section 3 gives a brief overview of the inductive version of the CP framework. Section 4 describes the proposed approach and defines the Non-conformity Measure (NCM) used for our ICP. Section 5 presents the experimental setting and performance measures used in our evaluation and reports the obtained experimental results. Finally, Section 6 gives our conclusions and plans for future work.

2. Related Work

EEG is a technique used to capture and record brain activity by measuring voltage fluctuations on different scalp areas according to the international electrode placement system. EEG captures the amplitude and the frequency of the electrical signal. The amplitude ranges from $5\mu V$ to $200\mu V$ and the frequency ranges from 0-100Hz (American-EEG-Society, 1994). It is the most common technique that has been utilized in BMI. It's popularity is due to several advantages: EEG signals are non-invasive, low cost, compatible, portable and have a high temporal resolution in comparison with other brainwave measurements such as electrocorticograms (ECoGs), magnetoencephalograms (MEGs), functional magnetic resonance imaging (fMRI) and near-infrared spectroscopy (fNIRS) (Al-Quraishi et al., 2018).

EEG's have several characteristic sub bands called brain waves. Brain waves are indicated by Greek letters: Delta 0 to 4 Hz, Theta 4 to 8 Hz, Alpha 8 to 12 Hz, Beta 12 to 32 Hz and Gamma 32 to 64 Hz.

Delta: Delta waves have a frequency range below 4Hz. They can be observed during sleep state and in infants, or in serious organic brain diseases. Irregular delta wave activity with a frontal emphasis is related to destructive or compressive lesions involving the diencephalon and upper midbrain, to deep frontal lesions, and to acute metabolite and

electrolyte disturbances. Animals are known to have more activity in this range (Thakor and Sherman, 2013).

Theta: Theta waves have a frequency range of 4-8Hz. In a child's brain they occur mainly in parietal and temporal regions. In healthy and alert adults, such theta wave activity is generally inconspicuous or absent, but it does appear during periods of disappointment, frustration, stress and certain stages of sleep as mentioned above (Gómez-Gil et al., 2014). It should be noted that theta activity appears after a generalized seizure, in patients with metabolic disorders, white matter encephalopathy, or extensive lesions of the upper brainstem (Thakor and Sherman, 2013).

Alpha: Alpha waves have a frequency range of 8-12Hz. These waves can be recorded from the occipital region (and sometimes from parietal and frontal regions as well) during consciousness, and is weakened by visual and other sensory stimulus. Alpha waves can be observed in a relaxed person and when eyes are closed. As mentioned above Alpha waves tend to disappear in sleeping subjects (Thakor and Sherman, 2013).

Beta: Beta waves have a frequency range of 12-32Hz. Beta waves can be recorded from the frontal and parietal lobes. Beta waves can be observed during intense mental activity and tension (American-EEG-Society, 1994).

Gamma: Gamma waves have a frequency range of 32-64Hz. Gamma waves are associated with brain processing. Studies using intra cranial electrodes showed that Gamma activity is related with states of high attention, conscious perception, information processing and motor activity. Note than when we record EEG's there exist three sources of artifacts that affect the Gamma Band :the 1st one is the power source noise 50/60Hz depending on the country, the 2nd one is EMG (Electromiograph) from the scalp and neck muscles and the 3rd one is electrical potentials produced by eye muscle contraction at the start of saccades (saccades are quick, simultaneous movements of both eyes in the same direction). This results to the need of developing effective filters to handle these artifacts (Nottage, 2009).

EOG is a technique that can be used to estimate eye orientation, it is based on the voltage amplitude modulation between two electrodes placed around the eyes. This voltage is depended by the angle of the eye. In humans the EOG signal ranges from 0.05 to 3.5 mV, with a typical bandwidth of DC to 50 Hz (López et al., 2016).

In the remaining of this section we provide some of the related work carried out. Witkowski et al. (2014) investigated if the use of EEG's fused with EOG's can enhance reliability and safety of continuous brain control of a hand exoskeleton performing grasping motions. Their results suggest that using different bio signals can increase usability and safety. Their dataset includes EEG and EOG's from 12 healthy right handed volunteers of age ranging from 25 to 32. The EEG's were recorded from 5 conventional EEG recording sites at a sampling rate of 200hz bandpass filtered at 0.4-70hz and preprocessed with a small Laplacian filter. The EOG's were recorded in accordance to the standard EOG placements at the left and right outer canthus (LOC/ROC). To evaluate brain-neural computer interaction (BNCI) control and safety, participants follow a visual cue indicating either to move or not to move the hand exoskeleton in a random order. Movements exceeding 25% of a full grasping motion when the device was not supposed to be moved were defined as safety violation. Participants reached comparable control but safety was frequently violated when using EEG, but not when using both EEG and EOG. Their discrimination

method was based on calculating the power spectrum of each EEG clip, selecting the frequency that showed largest event-related desynchronization (ERD) during motor imagery and event-related synchronization (ERS) during rest recorded from a C3 electrode. Based on the maximum values for ERD and ERS, a discrimination threshold was set. On using the EOG's a detection threshold for full left and right eye movements was set to reset the exoskeleton and enhance safety. When the visual cue was shown either green or red the true positive rate was $63.59 \pm 10.81\%$ and $60.77 \pm 9.42\%$ using EEG's and EEG's fused with EOG's respectively while the false positive rate was $36.11 \pm 10.85\%$ and $12.31 \pm 5.39\%$.

Gordleeva et al. (2017) proposed the development of a neuro-integrated control system for a lower-limb robotic exoskeleton (RE) using brain computer interface (BCI) technology based on recognition of EEG patterns evoked by motor imagery of limb movement. Their experiment involved 14 subjects (6 males and 8 females aged 18 to 23 years). While performing the experiment the exoskeleton was near the subject and performed movements with the right or the left foot depending on the EEG pattern generated by the subject. Their study consists of three main modules: EEG signal recording, EEG signal classifier and the software for transmission of commands to RE. On recording the signal 8 electrodes have been used at a sampling rate of 500 Hz filtered at 6-15hz with Notch filter of 50 Hz. The procedure of implementing the EEG signal classifier is as follows: The first two consecutive sessions where used for training and testing. These two sessions were used for initial setup and for testing the classifier. While recording the EEG's the subject followed the arrow orders that appeared on the screen: rest, left and right movement. When EEG recording ended, a spatial filtering was applied. Afterwards, the power in the frequency range from 6 to 15 Hz with a step of 1 Hz was calculated for each data set corresponding to the stimulus type in each channel separately. Then the rate of power change relative to the rest of the corresponding motor imagery was calculated. When spectral power decreased by more than 50% during motor imagery, the operator was considered to have successfully mastered the motor imagery technique and proceeded to the test session of the classifier. When several failed attempts occurred the procedure was repeated with the change of motor imagery type. While testing the classifier the subject was provided with feedback that showed if there was an agreement between the orders and the recognized task. The classification was based on discriminant analysis using the features identified by spatial filtering for type of orders pairwise, then the labels were classified using voting. After training and testing finished the subject was able to control the RE. The classifier analyzed EEG clips of 4.5 seconds and transmitted the appropriate command to the RE. Finally the authors report an average accuracy(over all subjects) of 73% 71% 66% for the three sessions.

Soekadar et al. (2014) used a similar approach with Witkowski et al. (2014) introducing a BNCI system that fuses electroencephalography (EEG) and electrooculography (EOG) to improve control of assistive robotics in daily life environments. The performance was evaluated on four men, with average age 26.5 ± 3.8 years and a 34-year-old patient with complete finger paralysis due to a brachial plexus injury. While recording EEG's they used 7 electrodes at a sampling rate of 200hz bandpass filtered at 0.4-70hz. When they used only EEG's they obtained an accuracy of $70.24 \pm 16.71\%$ for the healthy volunteers and an accuracy of $65.93 \pm 24.27\%$ for the patient, while when they incorporate EOG's they obtained an accuracy of 80.65 ± 11.28 and $76.03 \pm 18.32\%$ respectively.

Liu et al. (2017) implemented an EEG-based brain-controlled lower-limb exoskeleton for gait training. Their experiment involved six subjects aged from 21 to 26 years. No neurological or psychiatric disorders were reported among subjects. EEG signals were recorded from 16 positions using a sampling rate of 512 Hz. EEG signals were power ling notch filtered and band passed at 0.1 to 100 hz. Each subject performed three recording sessions the first one for the classifier training, the second one for on-line testing with visual feedback, and the last one for real-time robot control. Two asynchronous signal modalities have been used to classify the movement intention detection, sensorimotor rhythms (SMRs) and movement-related cortical potentials (MRCPs). They have obtained an accuracy of 80.16%±5.44% and 68.62±8.55% using SMR-based and MRCP-based methods respectively.

Even though the aforementioned studies produced promising results, still the obtained accuracy is rather low for such a safety critical application. This is the main motivation for our work, which tries to address this issue with the provision of confidence information that can be used for controlling the error rate of the technique and thus enhancing safety.

3. Inductive Conformal Prediction

For the needs of this study an ICP is built on top of the underlying model. The reason we used the Inductive version of CP and not the original Transductive version is because training deep Artificial Neural Networks is a time intensive task. In the remaining of this section we give a brief description of the main principles of inductive conformal prediction (ICP). For more details see Vovk et al. (2005).

Let $A = \{(x_i, y_i)|i = 1, \ldots, N\}$ denote our training set, where x_i is an object given in the form of an input vector or matrix, $R = \{t_1, \ldots, t_c\}$ is the set of possible labels and $y_i \in R$ is the label of the corresponding input vector or matrix. Let $C = \{(x_l, y_l)|l = 1, \ldots, L\}$ denote our calibration set and let $B = \{X_k|k = 1, \ldots, M\}$ denote our test set, where X_k is a test instance (vector or matrix). These sets will lead us to assessing predictions with confidence measures and finding which candidate labels are possible for the test instance X_k given a desired confidence level.

A *nonconformity* score (NCS) is a numerical value assigned to each instance that indicates how unusual or strange a pair (x_s, y_s) is, based on the underlying algorithm, where $s \in \{1, \ldots, L, new\}$ is the index of the sth element in C and new denotes a test example belonging to B. In particular, the underlying algorithm is trained on the instances belonging to A and the NCM uses the resulting model to assign a NCS denoted as $\alpha_{\tilde{s}}$ to each element belonging to C and a NCS denoted as $\alpha_{new}^{k,l}$ for every test instance in B. Using the calibration set denoted as C we calculate a sequence of NCS denoted as H containing L elements. For every test example k we use the sequence H to find the p-value of a test example k with a candidate label t_l. Given the sequence H and the NCS of a test example k we can calculate the p-value of a test instance (X_k, t_l) with the function:

$$p_k(t_l) = \frac{|\{\alpha_{\tilde{s}} \in H|\alpha_{\tilde{s}} \geq \alpha_{new}^{k,l}\}| + 1}{L + 1}, \tag{1}$$

where $\alpha_{new}^{k,l}$ is the NCS of the k^{th} example in the test set with candidate label t_l and $\alpha_{\tilde{s}}$ is the NCS of the s^{th} example in the calibration set.

Given a pair (X_k, t_l) with a p-value of δ this means that this example will be generated with at most δ frequency, under the assumption that the examples are exchangeable, proven in Vovk et al. (2005).

After all p-values have been calculated they can be used for producing prediction sets that satisfy a preset confidence level $1 - \delta$ (δ is called the significance level). Given the significance level δ, a CP will output the prediction set:

$$\{t_l | p_k(t_l) > \delta\}.$$

We would like prediction sets to be as small as possible. The size of prediction sets depends on the quality of the p-values and consequently on the NCM used.

If we want only a single prediction, or *forced prediction*, the CP outputs the label t_r with

$$r = \arg\max_{l=1,...,c} \; p_k(t_l),$$

in other words the t_l with the highest p-value. These predictions are complemented with measures of *confidence* and *credibility*. Confidence is defined as one minus the second largest p-value. Confidence is a measure that indicates the likelihood of a predicted classification compared to all the other possible classifications. Credibility is defined as the largest p-value. Low credibility means that either the data violate the exchangeability assumption or the particular test example is very different from the training set examples.

4. Proposed Approach

In this section we describe the proposed approach in terms of the data transformation we used as preprocessing, the Bidirectional Long Short Term Memory (BiLSTM) Neural Network Breiman (1996), which we used as underlying technique, and the NCM we used for the BiLSTM-ICP.

4.1. Discrete Fourier Transformation

Before fitting the BiLSTM Neural Network with EEG a Discrete Fourier Transformation (DFT) is applied to calculate the amplitude which will be the input of our Neural Network. The DFT is an algorithm which decomposes a sequence of values into components of different frequencies. It converts a finite list of equally spaced samples of a function into the list of coefficients of a finite combination of complex sinusoids, ordered by their frequencies, that have those same sample values:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N}, k = 0, ..., N-1, \tag{2}$$

where X_k is a complex number that encodes both amplitude and phase of a sinusoidal component of function x_n. The sinusoid's frequency is k/N cycles per sample. In our case X_k is a vector component of a vector produced from DWT. Its amplitude and phase are:

$$|X_k| = (\mathrm{Re}(X_k)^2 + \mathrm{Im}(X_k)^2)^{0.5}/N, \tag{3}$$

$$\mathrm{Arg}(X_k) = -i \cdot \ln(X_k/(|X_k|)). \tag{4}$$

4.2. Bidirectional Long Short Term Memory Neural Network

A BiLSTM is a combination of Bi-directional Recurrent Networks (Bi-RNN) and Long Short Term Memory(LSTM). A Recurrent Neural Network (RNN) is a Neural Network that is able to handle and extract information within the dependencies of sequences and time series of data. This is due to the fact that its connections form a directed graph along a sequence or a time series. However RNN can not handle long data sequences because of the vanishing and exploding problem of the gradient. To overcome the gradient problem and handle long data sequences the LSTM Neural Network was proposed where the gradient computation is truncated at certain architecture specific points, without affecting the long term error flow (Hochreiter and Schmidhuber, 1997). LSTM relies on a structure called a memory cell, which is composed of four main elements: an input gate, a neuron with a self recurrent connection, a forget gate and an output gate. On one hand both LSTM and RNN can get information from the previous context, on the other hand Bi-RNN can get information from the front and back. Thus a combination of Bi-RNN and LSTM should be beneficial because it utilizes the advantages of both LSTM and Bi-RNN (Yulita et al., 2017).

To improve accuracy, generalization and training speed the sequence input layer is connected to a fully connected layer of 150 neurons, which is then connected to a Rectified Linear Unit (ReLU) layer where each negative element of the input is set to zero. This is followed by a BiLSTM layer consisting of 150 neurons, the output of which is fed to a fully connected layer with a number of neurons equal to the number of classes. Finally a softmax layer provides probabilistic class outputs.

The BiLSTM Neural Network outputs class probabilities which are used to compute NCM's. All of our Neural Networks have been trained on an Nvidia RTX 2070 using the Adam optimizer for minimizing cross-entropy loss, with the number of epochs set to 50 and the learning rate to 0.0001.

4.3. Nonconformity measure

In this section we provide a description of the NCM used in this study. The NCM is based on the probabilities provided by the BiLSTM described in Section 4.2. Recall from Section 3 that $\{t_1, \ldots, t_c\}$ are the possible labels (note that we deal with a binary problem thus in our case c=2), $\alpha_{\tilde{s}}$ is the NCM of the sth instance (x_s, y_s) belonging to the calibration set and $\alpha_{new}^{k,l}$ is the NCM of the kth instance (X_k, t_l) in the test set. The nonconformity measure for a pair (x_s, y_s) or (X_k, t_l) respectively is:

$$\alpha_s = -D_{y_s}^s, \quad \alpha_{new}^{k,l} = -D_{t_l}^k, \tag{5}$$

where $D_{y_s}^s$ is the probability of prediction y_s for the calibration instance s and $D_{t_l}^k$ is the probability of prediction for candidate label t_l for the test instance k. I.e the higher the probability of class y_s and t_l the less nonconforming the examples. After calculating the NCS we calculate p-values and make predictions following the process described in Section 3.

5. Experiments and Results

5.1. Dataset

5.1.1. Experimental Paradigm

The dataset we used consists of EEG and EOG data collected from one subject with a high spinal cord lesion. The subject attempted to operate a neuroprosthetic device attached to his paralyzed right upper limb. The cue base BCI paradigm consisted of two tasks, the imagination of movement of the right hand and relaxation/no movement.

A visual signal randomly indicated to the user to either move or not move the exoskeleton. The two indications were displayed a total of 24 times for 5 seconds and then the exoskeleton was reset to the open position. Each indication is separated by inter-trial intervals of 4-6 seconds. The above process was repeated three times. Specifically the dataset consists of three sets each one stored in a Matlab structure. The data file is available at http://bnci-horizon-2020.eu/database/data-sets.

5.1.2. Data Acquisition

EEG was recorded at a sampling rate of 200Hz using 5 conventional EEG recording sides, bandpass filtered at 0.4-70Hz, while EOG was recorded at a sampling rate of 200Hz in accordance to the standard EOG placements at the left and right outer canthus (LOC/ROC).

5.2. Experimental Setting and Performance Measures

Here we detail the experiments and results of the proposed approach. As we mentioned in Section 1 for the needs of this study we have trained two BiLSTM Neural Networks. The first one is fitted with the amplitude of the EEG clips while the second one is fitted with the amplitude of the EEG clips together with the EOG clips. For calculating the amplitude of the EEG's we first applied a DFT. Note that the EOG clips when fitted are fitted intact.

We used the three sets as training, calibration and test sets. More specifically we used each one as test set performed in total six runs to obtain results.

First we report forced prediction results in terms of accuracy and mean confidence and credibility. Due to the fact that the accuracy itself is not a good indication for measuring the performance of a CP we used four probabilistic criteria for evaluating p-values proposed by Vovk et al. (2016). These criteria are divided into two main categories called *Basic Criteria*, which do not take into account the true label, and *Observed Criteria*, which take into account the true label. The two Basic Criteria we used are:

The S ("sum") criterion

$$\frac{1}{M}\sum_{l=1}^{c}\sum_{k=1}^{M} p_k(t_l), \tag{6}$$

where $p_k(t_l)$ is the p-value of the test example X_k with candidate label t_l as in equation (1) In effect the S-criterion is the average sum of all p-values.

The N ("number") criterion

$$\frac{1}{M}\sum_{k=1}^{M} |\{t_l | p_k(t_l) > \delta\}|, \tag{7}$$

which is the average size of the prediction sets with respect to a confidence level $1-\delta$.

The two Observed Criteria we used are:

The OF ("observed fuzziness") criterion

$$\frac{1}{M}\sum_{k=1}^{M}\sum_{l,t_l\neq t_k} p_k(t_l), \tag{8}$$

which uses the average sum of the p-values of the false labels.

The OE ("observed excess") criterion

$$\frac{1}{M}\sum_{k=1}^{M}|\{t_l|p_k(t_l)>\delta, t_l\neq t_k\}|, \tag{9}$$

which represents the average number of false labels included in the prediction sets, with respect to a confidence level $1-\delta$.

For all criteria smaller values indicate more informative p-values.

5.3. EEG Classification results

5.3.1. Accuracy

Table 1 presents the maximum accuracy obtained of the conventional BiLSTM techniques when using as input the EEG amplitude and when using the EEG amplitude together with the EOG clips. We can observe that when we use only the EEG amplitude the resulting accuracy is higher than when we use the EEG amplitude together with the EOG clips. This is due to the fact that the EOG clips are noisy and non stationary thus the BiLSTM technique could not extract additional useful information for the classification task.

Table 2 reports the accuracy of the corresponding Conformal Prediction technique along with the average confidence and credibility measures it produced. The accuracies reported in these tables are quite high for all problems and comparable to that of the related work presented in Section 2. When comparing the accuracy of the Inductive Conformal Predictor technique with that of its conventional counterpart, we observe that the first leads to a very small decrease. However this decrease is not significant especially since our main interest is on the prediction sets provided by CP.

Table 1: Accuracy of the Underlying models

	Accuracy(%)
EEG	74.80
EEG+EOG	72.92

5.3.2. Empirical Validity

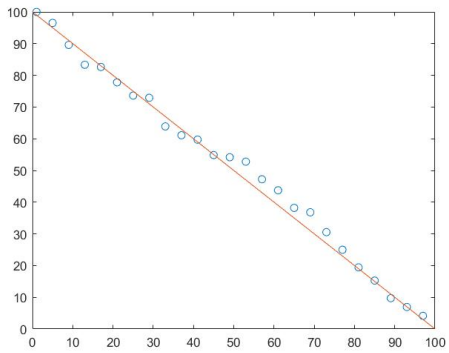
Figure 1, presents the percentage of correct region predictions(the percentage of the number of the sets that contain the correct class) as a function of the significance level for the two BiLSTM-ICPs. In all cases the accuracy is almost equal to the required confidence level, i.e.

Table 2: Average accuracy, credibility and confidence of the proposed CP's

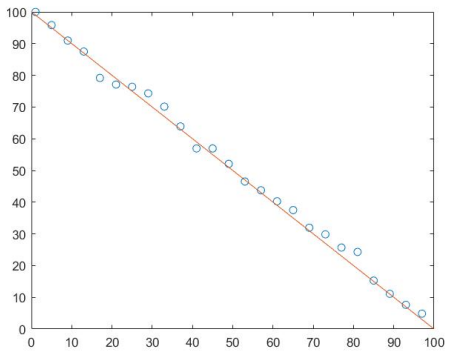
Accuracy	EEG	74.31
	EEG+EOG	70.14
Average confidence	EEG	84.50
	EEG+EOG	84.03
Average credibility	EEG	63.50
	EEG+EOG	63.97

the observations are close to the diagonal. We can observe some decline from the diagonal, in fact sometimes the accuracy is smaller that the confidence level, this is due to the difficult nature of correctly classifying EGG's and EOG's which is non stationary and noisy data. Another reason is statistical fluctuations as we used 3 fixed sets as training calibration and test sets, without performing any form of randomization, each one containing 24 instances.

Figure 1: Percentage of correct region predictions with respect to the significance level using EEG and EEG+EOG as inputs



(a) EEG



(b) EEG+EOG

5.3.3. Informational Efficiency

In this study the main objective is to complement the single predictions with probabilistic measures of confidence and provide prediction sets with respect to a confidence level. Here we investigate how informative our p-values are and the practical usefulness of our prediction sets. This is done following the informational efficiency criteria described in Subsection 5.2 and proposed by Vovk et al. (2016).

Table 3 shows the values of the two unobserved criteria for the BiLSTM-ICP classifier using as inputs EEG's or EEGs+EOGs. The second column of the table contains the values of the S criterion, while the rest of the columns present the N criterion for the significance levels 0.01, 0.05, 0.10 and 0.20. Table 4 presents the values of the two observed criteria. The second column contains the values of the OF criterion, while the rest of the columns

Table 3: Unobserved criteria

Classification Problem	S criterion	N criterion (per significance level) 0.01	0.05	0.10	0.2
EEG	0.79	2	1.7431	1.5833	1.2083
EEG+EOG	0.7994	2	1.7292	1.6042	1.2014

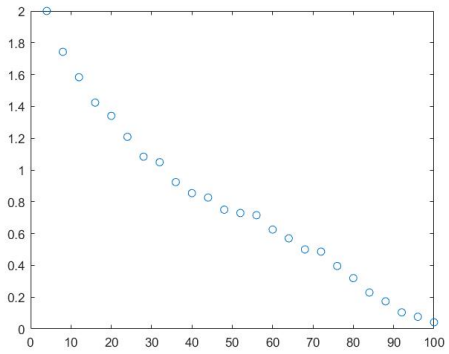
Table 4: Observed criteria

Classification Problem	OF criterion	OE criterion (per significance level) 0.01	0.05	0.10	0.20
EEG	0.27	1	0.78	0.6875	0.4306
EEG+EOG	0.2794	1	0.77	0.6944	0.4306

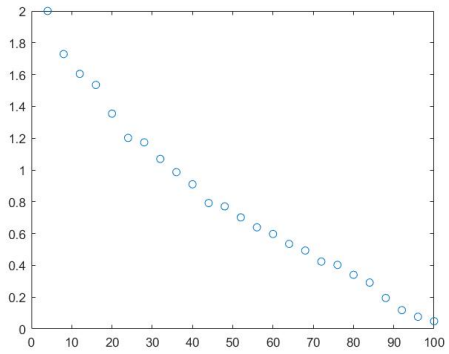
give the values of the OE criterion for the significance levels 0.01, 0.05, 0.10 and 0.20. The values of all criteria show how informative our p-values are, furthermore the values of the N and OE criteria also demonstrate the practical usefulness of the produced prediction sets.

When we set the significance level to 0.1 the prediction sets contain on average approximately 1.6 labels, while the number of false labels is 0.69. When we increase the significance level to 0.2 the average size contain approximately 1.2 labels and the number of false labels is 0.43. Both techniques have similar values of S and OF criterion indicating that there is no significant difference in the quality of the p-values. A graphical illustration of the size of the prediction sets and the number of false labels as a function of the significance level can be seen in figure 2 and figure 3 respectively.

Figure 2: Average size of region predictions with respect to the significance level using EEG and EEG+EOG as inputs

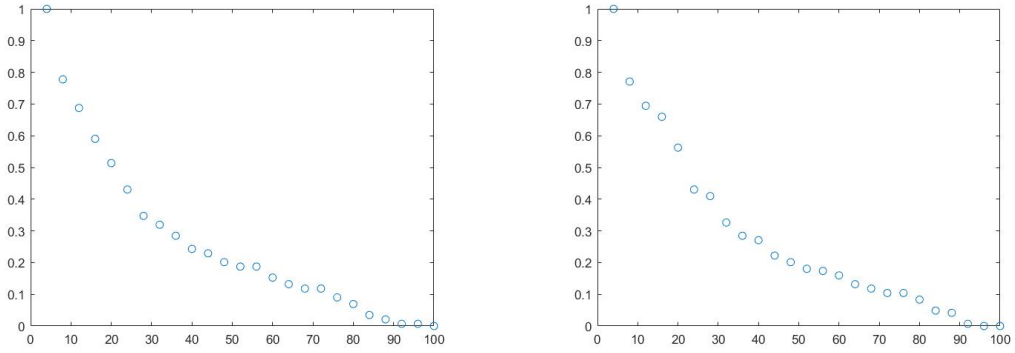


(*a*) EEG



(*b*) EEG+EOG

Figure 3: Average number of wrong labels contained in region predictions with respect to the significance level using EEG and EEG+EOG as inputs



(a) EEG (b) EEG+EOG

6. Conclusions

We examined the use of Conformal Prediction for providing confidence measures to assist a BMI in controlling an exoskeleton using EEG and EOG clips. More specifically we developed an ICP based on a Bidirectional Long Short Term Memory (BiLSTM) Neural Network and examined it's performance on a dataset consisting of EEG and EOG data collected from one subject with a high spinal cord lesion. The main idea behind this study is addressing the serious safety risks of unwanted actions that can be caused by misclassifications. CP can be used for controlling the frequency of misclassifications and postponing any actions for which there is uncertainty until more evidence towards the right action is received and a confident enough decision can be made.

We evaluated the performance of the proposed ICP approach using EEG's and the combination of EEG's with EOG's. Our results indicate that the obtained accuracy is comparable to that of the other techniques existing in the literature for the particular task, while providing important confidence information that can enhance safety. The produced prediction sets are well calibrated, which shows that the proposed approach can successfully control the error rate of its outputs. Furthermore, considering the relatively low accuracy of existing conventional techniques on the particular task, the quality of the obtained p-values and the tightness of the corresponding prediction sets are arguably a good result. We believe that in a real scenario, reinforced signal leading to a certain prediction of the desired action can be given very fast, therefore uncertainties will only cause minor delays. We would like to verify this in our future work.

Additionally, the dataset we used in this study is relatively small, so another future plan is to evaluate the performance of the proposed approach on larger datasets including more subjects. Finally, we would also like to investigate the use of other deep ANN structures or ensemble techniques as underlying algorithms, which we believe will result in performance improvements.

References

Maged Al-Quraishi, Irraivan Elamvazuthi, Siti Asmah Daud, S Parasuraman, and Alberto Borboni. Eeg-based control for upper and lower limb exoskeletons and prostheses: A systematic review. *Sensors*, 18, 10 2018. doi: 10.3390/s18103342.

American-EEG-Society. Guideline fifteen: guidelines for polygraphic assessment of sleep-related disorders (polysomnography). american electroencephalographic society. *Journal of clinical neurophysiology : official publication of the American Electroencephalographic Society*, 11 1:116–24, 1994.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, Aug 1996. ISSN 1573-0565. doi: 10.1007/BF00058655. URL https://doi.org/10.1007/BF00058655.

Denise L Brown-Triolo, Mary Joan Roach, Kristine Ann Nelson, and Ronald J. Triolo. Consumer perspectives on mobility: implications for neuroprosthesis design. *Journal of rehabilitation research and development*, 39 6:659–69, 2002.

Pilar Gómez-Gil, Ever Juárez-Guerra, Vicente Alarcón-Aquino, Manuel Ramírez-Cortés, and José Rangel-Magdaleno. *Identification of Epilepsy Seizures Using Multi-resolution Analysis and Artificial Neural Networks*, pages 337–351. Springer International Publishing, Cham, 2014. ISBN 978-3-319-05170-3. doi: 10.1007/978-3-319-05170-3_23. URL https://doi.org/10.1007/978-3-319-05170-3_23.

Susan Gordleeva, M.V. Lukoyanov, Sergey Mineev, M.A. Khoruzhko, Vasily Mironov, Alexander Kaplan, and Victor Kazantsev. Exoskeleton control system based on motor-imaginary brain–computer interface. *Sovremennye tehnologii v medicine*, 9:31, 09 2017. doi: 10.17691/stm2017.9.3.04.

Hui Fan He and Kazuo Kiguchi. A study on emg-based control of exoskeleton robots for human lower-limb motion assist. *2007 6th International Special Topic Conference on Information Technology Applications in Biomedicine*, pages 292–295, 2007.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9 (8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Dong Liu, Weihai Chen, Zhongcai Pei, and Jianhua Wang. A brain-controlled lower-limb exoskeleton for human gait training. *Review of Scientific Instruments*, 88:104302, 10 2017. doi: 10.1063/1.5006461.

A. López, F. J. Ferrero, M. Valledor, J. C. Campo, and O. Postolache. A study on electrode placement in eog systems for medical applications. In *2016 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–5, May 2016. doi: 10.1109/MeMeA.2016.7533703.

Judith Nottage. Uncovering gamma in visual tasks. 23:58–71, 12 2009.

Surjo Soekadar, Matthias Witkowski, Nicola Vitiello, and Niels Birbaumer. An eeg/eog-based hybrid brain-neural computer interaction (bnci) system to control an exoskeleton for the paralyzed hand. *Biomedizinische Technik. Biomedical engineering*, 60, 12 2014. doi: 10.1515/bmt-2014-0126.

Nitish V. Thakor and David L. Sherman. *EEG Signal Processing: Theory and Applications*, pages 259–303. Springer US, Boston, MA, 2013. ISBN 978-1-4614-5227-0. doi: 10.1007/978-1-4614-5227-0_5. URL https://doi.org/10.1007/978-1-4614-5227-0_5.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer Science & Business Media, 2005.

Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alexander Gammerman. Criteria of efficiency for conformal prediction. pages 23–39, 2016. doi: 10.1007/978-3-319-33395-3_2. URL http://dx.doi.org/10.1007/978-3-319-33395-3_2.

Matthias Witkowski, Mario Cortese, Marco Cempini, Jürgen Mellinger, Nicola Vitiello, and Surjo R. Soekadar. Enhancing brain-machine interface (bmi) control of a hand exoskeleton using electrooculography (eog). *Journal of NeuroEngineering and Rehabilitation*, 11 (1):165, Dec 2014. doi: 10.1186/1743-0003-11-165. URL https://doi.org/10.1186/1743-0003-11-165.

Intan Nurma Yulita, Mohamad Ivan Fanany, and Aniati Murni Arymuthy. Bi-directional long short-term memory using quantized data of deep belief networks for sleep stage classification. *Procedia Comput. Sci.*, 116(C):530–538, November 2017. ISSN 1877-0509. doi: 10.1016/j.procs.2017.10.042. URL https://doi.org/10.1016/j.procs.2017.10.042.