Extracting Human Levels of Trust in Human-Swarm Interaction using EEG signals

Jesus A. Orozco and Panagiotis Artemiadis*, *IEEE Senior Member*

***Abstract*—Trust is an essential building block of human civilization. However, when it relates to artificial systems, it has been a barrier to intelligent technology adoption in general. This paper addresses the gap in determining levels of trust in scenarios that include humans interacting with a swarm of robots. Electroencephalography (EEG) recordings of the human observers of the different swarms allow for extracting specific EEG features related to different trust levels. Feature selection and machine learning methods comprise a classification system that would allow recognition of different levels of human trust in those human-swarm interaction scenarios. The results of this study suggest that EEG correlates of swarm trust exist and are distinguishable in machine learning feature classification with very high accuracy. Moreover, comparing common EEG features across all human subjects used in this study allows for the generalization of the classification method, providing solid evidence of specific areas and features of the human brain where activations are related to levels of human-swarm trust. This work has direct implications for effective human-machine teaming with applications to many fields such as exploration, search and rescue operations, surveillance, environmental monitoring, and defense. In those applications, quantifying levels of human trust in the deployed swarm is of utmost importance because it can lead to swarm controllers that adapt their output based on the human's perceived trust level.**

***Index Terms*—Electroencephalography, human-swarm interaction, trust.**

I. Introduction

Trust is an essential building block of human civilization. Trusting others allows us to create and maintain relationships, and trusting systems, such as governments, allow us to create and maintain societies [1]. When it comes to artificial systems, people and governments become weary, delaying artificial intelligence technology adoption due to a lack of trust in the intelligent systems [2], [3]. Unlike humans, robots can not sense and subsequently adapt to a human's growing distrust of their actions. A framework to quantify trust between humans and robots could enhance human-machine teaming across a variety of applications. However, such a framework is still missing.

While many definitions exist, in this work, we adopt the following definition of trust used in human-autonomy interaction: "The attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [4]. One specific area that we explore trust is within Human-Swarm Interaction (HSI). A swarm is a group of robot agents with basic control laws, such as moving away from each other, moving towards one another, or matching another neighbor's heading; a combination of these control laws allows for an organic movement of the robots, similar to a flock of birds or school of fish [5]. Previous work on human-swarm interaction shows that collective swarming behaviors are distinguishable at the brain level [6]; a necessary step for an HSI brain-computer interface. Although there are many methods to teleoperate a swarm, [7]–[10], in order to take the leap in automating aspects of this control, we need to trust that the swarm will do whatever task is assigned to it, or ensure that the swarm can sense and adapt to our growing distrust in their actions.

While the research focused on trust in HSI is sparse, trust is investigated in human-machine interactions, with research focused on predictive models (e.g., convolutional neural network, Markov decision process) to establish different levels of trust [11]–[13]. These works establish that there are many factors (e.g., demographics, status information, dependability, transparency) that influence trust, but trust can be modeled adequately if these factors are considered. As trust and control of swarms are usually intertwined for effective human-swarm teaming, researchers have used predictive models of trust to control a swarm in the past [14]. This research showed that trust is affected by factors such as the swarm's physical appearance and the user's conceptualization of trust, which resulted in each study participant requiring their own trust model [14]. However, the model described in this study only captures intervention and non-intervention features as a trust metric. A more robust model would benefit from other factors that might give insight into human intent; however, this also requires user input of their subjective trust ratings to provide trust feedback effectively [14].

Indications of trust have been observed at the brain level before, via Magnetic Resonance Imaging (MRI) studies [15], as well as ElectroEncephaloGraphy (EEG) [16]–[20]. It has been shown that different brain regions contribute to trust in human-autonomy interaction [17]. Specifically, trust in human-autonomy systems originates in the left-frontal, fronto-central, right-frontal, and occipital regions [17]. Similarly, machine learning has been used to classify trust levels in human-machine interaction in the past, with a prediction accuracy of approximately 72% [16]. While current literature on human-autonomy interaction gives us insight into trust via EEG data, findings fail to determine EEG features of trust in human-swarm interaction scenarios. These features are necessary for

*This material is based upon work supported by the National Science Foundation under Grant No. #2014264, as well as by the U.S. Air Force Office of Scientific Research (AFOSR) award FA9550-18-1-0221.

Jesus A. Orozco and Panagiotis Artemiadis are with the Mechanical Engineering Department, at the University of Delaware, Newark, DE 19716, USA. jaon@udel.edu, partem@udel.edu

*Corrresponding author: partem@udel.edu

a more hands-on approach to intervening with misbehaving or adversarial swarms, among others.

This paper addresses the gap in determining levels of trust in human-swarm interaction by cycling through different swarming behaviors and using machine learning to determine which EEG features are related to trust. Adopting the definition of trust described earlier [4], we develop an experimental scenario in which specific swarming behaviors allow for a clear distinction between which swarms could complete a task and which could not. By doing so, we systematically allow for the subject's attitudes towards the swarms to align with those of low trust and high trust. EEG recordings of the human observers of the different swarms allow for extracting specific EEG features related to different trust levels. The methods in our study aid our EEG feature selection and classification to show that we can decide between low trust and high trust for each subject with high prediction accuracy. Moreover, comparing common EEG features across subjects allows for the generalization of the classification method to all subjects, providing solid evidence of specific areas and features of the human brain where activations are related to levels of human-swarm trust. The proposed work introduces EEG-based feature classification related to human trust in robotic swarms for the first time, which has direct implications for effective human-machine teaming with applications to many fields such as exploration, search and rescue operations, surveillance, environmental monitoring, and defense.

II. Methods

This work focuses on determining trust features in EEG recordings that would allow us to improve human-swarm interaction and interfaces. The goal is to develop a classification system that would allow recognition of different levels of human trust in human-swarm interaction scenarios. Below, we first introduce the data type and swarm simulator with the respective swarming behaviors and task descriptions. We then discuss the EEG processing workflow, including the feature selection and classification, used to determine trust levels.

A. Experimental Setup

EEG and gaze tracking data were collected for this study, but for the purposes of this paper, only EEG recordings are analyzed and discussed further. The EEG data were collected using a Brain Products ActiCHamp amplifier with 32 active electrodes. The active electrodes were placed on the subject's scalp using an ActiCAP head cap based on the International 10-20 system [21]. The data were recorded at 500 Hz. Figure 1 shows the experimental setup.

B. Task Description

The experiment is comprised of virtual non-holonomic agents displayed as discs with lines displaying their current heading tasked with reaching a target point on the computer screen. All agents are modeled as double integrators, driven by two-dimensional forces which directly affect the two-dimensional acceleration of the agents, similarly to [22].

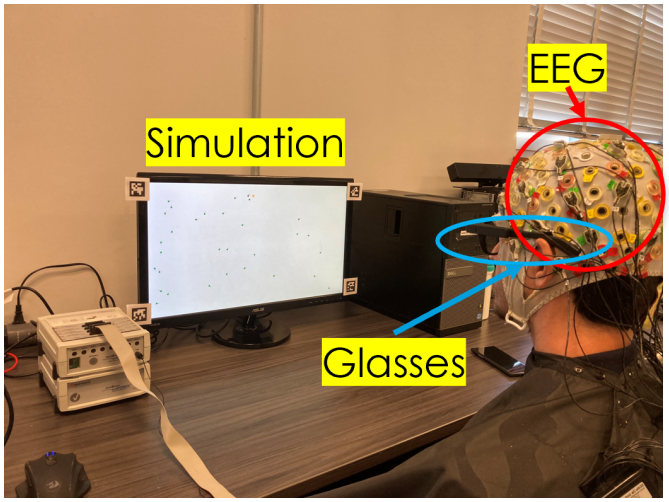

Fig. 1: Experimental setup. A subject is wearing an EEG cap and gaze-tracking glasses while viewing the computer screen during a swarm simulation.

There are three possible areas the swarm spawns from at the beginning of each simulation trial, with each agent having a random initial velocity and heading. Additionally, the swarm is presented with one of three different locations they attempt to reach (target), located at the top of the computer screen. These are as well randomized. To distinguish each swarm in post-experiment questions to the subject, we color-coded them depending on which behavior they exhibit at the time: green is *explore*, red is *surround*, and blue is *flock*. Based on previous work [23], each agent showed a line protruding from the perimeter of the agent indicating its current heading. Once an agent reaches the goal, it goes offline (i.e., it stops moving and its color turns gray). The task is complete when approximately 90% of the swarm of agents reaches the goal. If the swarm is unable to reach the goal under the allotted time of 25 seconds, then the simulation ends with the swarm failing to reach the goal. To note, the explore swarm never completes the task, while the flock and surround behaviors do in the allotted time. This is by design as we want to have a clear-cut definition for low (explore) and high trust (flock and surround) scenarios. Each subject observes each swarming behavior (*explore*, *surround*, and *flock*) five times for a total of 15 trials per subject. There is only one swarm exhibiting one behavior for the duration of a trial. Figure 2 shows an example of what subjects see during the experiment for one trial.

C. Swarm Simulator

The simulation of virtual agents displayed on the computer screen was generated and controlled using a Python 3.9.10 script and the *tkinter* package. The Lab Streaming Layer (LSL) Recorder software is used to time–synchronize both the EEG and gaze tracking data [24]. The simulation interface enables control of three different swarming behaviors by the experimenter while the simulation updates with new agent positions every 16 ms, or approximately at 60 Hz frequency. The simulation framework was inspired by the work in [25].

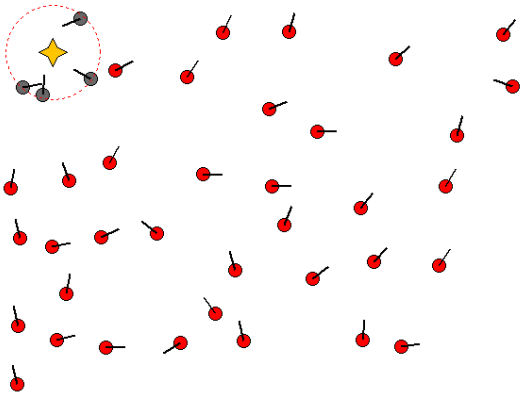

Fig. 2: Representative snapshot of the swarm simulation, where the agents are represented as discs with lines displaying their current heading. A red swarm is moving toward a goal shown as a yellow asterisk at the top left, with a few agents disabled (in gray) after reaching the goal's boundary, as indicated by a red dashed circle around the target.

D. Swarming Behaviors

The swarming behaviors are based on the work presented in [5], [26], with each behavior having a combination of repel, cohere, and align forces. We added a wall-bounce force to ensure the swarm stayed within the boundary of the computer screen. The control laws are applied differently depending on which swarming behavior is exhibited.

1) Explore: The *explore* swarming behavior is described by two force vectors, the repulsion vector, $\mathbf{c}_{r_i}$, and the wall-bounce vector, $\mathbf{c}_{wb_i}$. The force vector $\mathbf{c}_{r_i}$ for each agent i, $i = 1, 2, ..., n$, is computed as:

$$\mathbf{c}_{r_i} = \begin{cases} \sum_{j=1}^{n} (\mathbf{p}_i - \mathbf{p}_j), \text{ if } \|\mathbf{p}_i - \mathbf{p}_j\| \leq s_r, \text{ and } i \neq j \\ \mathbf{0}, \text{ if } \|\mathbf{p}_i - \mathbf{p}_j\| > s_r \end{cases} \tag{1}$$

where n is the swarm size, $\mathbf{p}_i$ is the position of the agent i, $\mathbf{p}_j$ is the position of the j^{th} neighbor, s_r is the sensing radius for the repulsion vector, and $\|\cdot\|$ denotes the vector norm.

When an agent reaches the wall perimeter of the simulation space, a wall-bounce force is exerted. The wall-bounce force $\mathbf{c}_{wb_i}$, which bounds the swarm to move within the screen, is computed by:

$$\mathbf{c}_{wb_i} = \begin{cases} \dfrac{\mathbf{w}_l - \mathbf{p}_i}{\|\mathbf{w}_l - \mathbf{p}_i\|^2}, & \text{if } \|\mathbf{w}_l - \mathbf{p}_i\| < s_{wb} \\ \dfrac{\mathbf{w}_r - \mathbf{p}_i}{\|\mathbf{w}_r - \mathbf{p}_i\|^2}, & \text{if } \|\mathbf{w}_r - \mathbf{p}_i\| < s_{wb} \\ \dfrac{\mathbf{w}_t - \mathbf{p}_i}{\|\mathbf{w}_t - \mathbf{p}_i\|^2}, & \text{if } \|\mathbf{w}_t - \mathbf{p}_i\| < s_{wb} \\ \dfrac{\mathbf{w}_b - \mathbf{p}_i}{\|\mathbf{w}_b - \mathbf{p}_i\|^2}, & \text{if } \|\mathbf{w}_b - \mathbf{p}_i\| < s_{wb} \end{cases} \tag{2}$$

with $\mathbf{w}_l$, $\mathbf{w}_r$, $\mathbf{w}_t$, and $\mathbf{w}_b$ defined as the normal vector of the agent to the left, right, top, and bottom screen boundaries, respectively. s_{wb} is defined as the sensing radius for the screen boundary. For the exploring behavior, s_r is set to 50 px, and s_{wb} is set to 60 px. The total size of the allowable simulation space for the agents is 1920 by 1080 px.

The total force exerted on each agent exhibiting the *explore* behavior is given by:

$$\mathbf{f}_{e_i} = \mathbf{c}_{r_i} + \mathbf{c}_{wb_i}. \tag{3}$$

2) Flock: The *flock* behavior makes use of repulsion and wall-bounce forces, similar to the *explore* behavior, but with different sensing radii for each. The sensing radii for $\mathbf{c}_{r_i}$ and $\mathbf{c}_{wb_i}$ under flock are set to 15 px and 50 px, respectively. In addition to these vectors, *flock* also incorporates an alignment force, $\mathbf{a}_{f_i}$, and a cohesion force, $\mathbf{c}_{cf_i}$. The $\mathbf{a}_{f_i}$ is different depending on the agent's leader status and computed as:

$$\mathbf{a}_{f_i} = \begin{cases} \mathbf{g}_g - \mathbf{p}_i, \text{ if } \|\mathbf{g}_g - \mathbf{p}_i\| < s_a \text{ and } l_{f_i} = true \\ \mathbf{a}_{f_{nl}}, \text{ if } \|\mathbf{g}_g - \mathbf{p}_i\| < s_a \text{ and } l_{f_i} = false \end{cases} \tag{4}$$

where $\mathbf{g}_g$ is the swarm global goal on the screen, $\mathbf{a}_{f_{nl}}$ is the alignment force of the nearest leader to the agent, s_a is the sensing radius of the alignment force, and l_{f_i} is the leader flag for the agent (i.e., set to $false$ if the agent is not a leader and $true$ is the agent is a leader). For this study, s_a is set to 100 px. If any agent is not a leader nor near one, then the agent continues along their current path. The $\mathbf{c}_{cf_i}$ force is defined as:

$$\mathbf{c}_{cf_i} = \mathbf{p}_i - a_r \mathbf{r}_f, \quad \text{if } \|\mathbf{p}_j - \mathbf{p}_i\| < s_{cf} \tag{5}$$

where s_{cf} is the sensing radius for the cohesion force under the *flock* behavior, a_r is defined as the agent radius, $\mathbf{r}_f$ represents a radius factor, in (x, y) coordinates, that acts as a boundary to avoid neighbor-agent collisions. For this study, s_{cf} is set to 50 px, a_r to 7 px and $\mathbf{r}_f$ to $(4, 4)$ px. As such, the total force exerted on each agent exhibiting the *flock* behavior is given by:

$$\mathbf{f}_{f_i} = \begin{cases} \mathbf{a}_{f_i} + c_w \mathbf{c}_{cf_i} + r_w \mathbf{c}_{r_i}, & \text{if } l_f = true \\ \mathbf{a}_{f_i} + c_w \mathbf{c}_{cf_i} + r_w \mathbf{c}_{r_i} + \mathbf{c}_{wb_i}, & \text{if } l_f = false \end{cases} \tag{6}$$

where c_w and r_w are the cohesion and repulsion weighting factors, respectively. For this study, if the agent is a leader, c_w is set to 0.9 and r_w to 1.0. If the agent is not a leader, then c_w is 0.9 and r_w is 2.5. Each swarm has 15 leaders that are randomly assigned at the beginning of each trial.

3) Surround: The *surround* swarming behavior is comprised of $\mathbf{c}_{r_i}$, $\mathbf{c}_{wb_i}$, and a modifed cohesion force, $\mathbf{c}_{cs_i}$. The $\mathbf{c}_{cs_i}$ force is defined as:

$$\mathbf{c}_{cs_i} = \mathbf{p}_i - mid(R), \text{ if } \|\mathbf{p}_j - \mathbf{p}_i\| < s_{cs} \tag{7}$$

where s_{cs}, set to 20 px, is the sensing radius for the *surround* behavior and $mid(\cdot)$ is the midpoint of a rectangle, R, defined with vertices $(x_{min}, y_{min}, x_{max}, y_{max})$. Figure 3 describes how R is obtained. The total force exerted on each agent exhibiting the *surround* behavior is given by:

$$\mathbf{f}_{s_i} = \mathbf{c}_{cs_i} + \mathbf{c}_{r_i} + \mathbf{c}_{wb_i}. \tag{8}$$

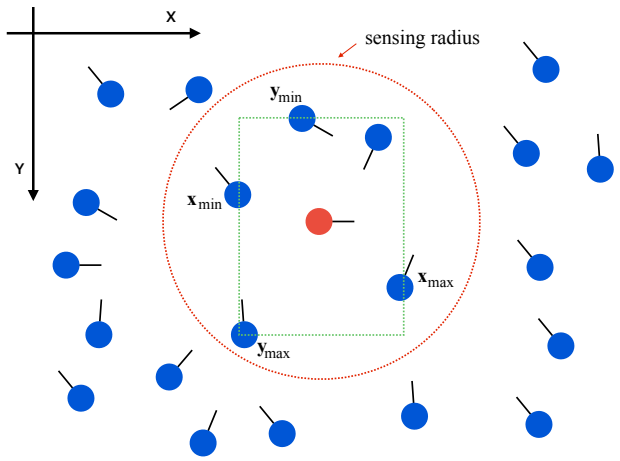

Fig. 3: The definition of the surround rectangle R used in the *surround* behavior is shown. The sensing radius concentric to the red agent, highlighted red only for visualization purposes, encapsulates five neighboring agents. R is constructed from the local minimum and maximum coordinates as shown by the x_{min}, y_{min}, x_{max}, and y_{max} labels.

E. Experiment Details

Criteria to participate as a human subject in this study are as follows: be between 21 – 50 years old, have no current or past history of visual impairment, have no known neurological issues, be in general good health, and be proficient in English. There was a total of seven subjects, two females and five males, that met these criteria and participated in the study with an average age of 25.4 $\pm$ 1.6 years. Due to a hardware malfunction, only six subjects had analyzable data. However, the number of subjects is in line with previous work related to human-robot interaction with the use of EEG [27]–[31]. The protocol for this study was approved by the University of Delaware Institutional Review Board (IRB ID: 1487701-5), with informed consent given by all subject participants.

Before the start of the experiment, we informed the subjects what they were going to see on the screen. Each subject observes each swarming behavior (*explore*, *surround*, and *flock*) five times for a total of 15 trials per subject. The subject is asked to think about each swarm's ability to reach the goal through each trial. Questions that they should have in mind are: "Is one swarm better equipped to handle this task?" In other words, "are you able to trust the swarm to complete the task?", or "is the swarm untrustworthy?" No verbal communication is required or expected by the subject during the trials. The subject's trust is therefore assessed through the demonstration of competence of the swarm to reach the goal, in line with how trust can be established [4].

After the experiment, each subject was asked about their impressions of the experiment based on the pre-experiment questions. Additionally, we asked the subjects if their answers would be different if the task was different. Specifically, they were asked: "If the task was to map out an area instead of reaching a specific point in space, which swarm would you choose and why?" All those questions were only for qualitative understanding of the human-swarm interaction scenario and they were not used in any quantifiable extraction of human trust later in the methods.

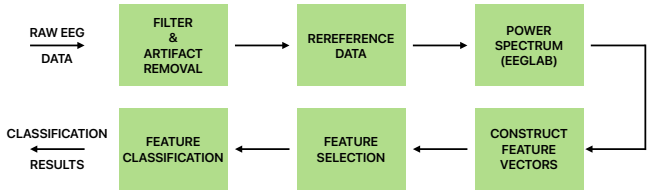

Fig. 4: EEG processing workflow.

F. Data Analysis

All the data are processed and analyzed in MATLAB and with the MATLAB plugin, EEGLAB, [32]. The workflow is detailed in Fig. 4. All feature selection and classification is done per subject since each individual experiences trust differently (see section I. Introduction for details).

1) EEG Data Processing: We process and clean up the data by applying a fourth-order low-pass Butterworth filter (cut-off frequency at 40 Hz), followed by a fourth-order high-pass Butterworth filter (cut-off frequency at 1 Hz), for removal of high-frequency noise and low-frequency trends, respectively. After filtering, we re-reference the data as an average of all the channels. Additionally, an electrooculogram (EOG) artifact removal algorithm [33] was applied to the processed EEG data to remove any eye movement or blinking artifacts. Finally, we created study sets in EEGLAB to extract power spectrum calculations for feature vector construction.

2) Feature Selection & Classification: After extracting all power spectrum values in the frequency domain from EEGLAB, we divided the data into five wavebands (0.5 Hz $< \delta <$ 4.0 Hz, 4 Hz $< \theta <$ 7.5 Hz, 8.0 Hz $< \alpha <$ 13.0 Hz, 14 Hz $< \beta <$ 26 Hz, 30 $< \gamma <$ 45.0 Hz) indicating the standard brain wave bands as described in [34]. After this step, we could average across the entirety of a waveband to obtain one feature per waveband per EEG channel. A concern with this approach is that it could mask a discernable feature if it were present in either the low or high end of the specific waveband. We, therefore, average the power spectrum values across sub-bands by splitting the band into approximately two equal-in-size sections to obtain a total of 10 power spectrum values, per channel. To illustrate this, Fig. 5 visualizes the average power spectrum density for one EEG channel, for low-trust and high-trust conditions; the vertical lines delineate the wave band barriers. As can be seen in Fig. 5, an average power spectrum value across the entirety of the beta band could make it so that this feature could not be used to distinguish between these two conditions. However, if each of the five brain wave bands (delta (δ), theta (θ), alpha (α), beta (β), gamma (γ)) is split into a low and high wave band, see Table I, it could have a significant difference in either sub-band for the feature to be used in classification. As such, there are a total of 10 features per EEG channel, for a total of 320 features per trial.

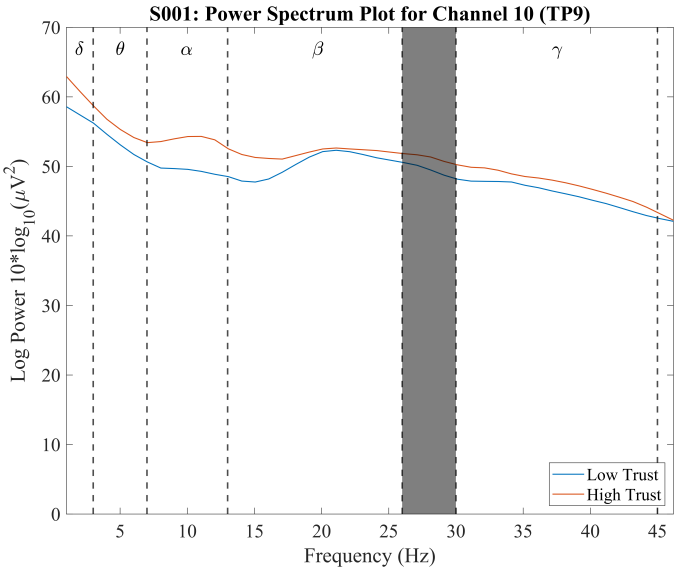

Fig. 5: An indicative power spectrum plot for channel 10, of subject 1, for one trial. The vertical dashed lines indicate the brain wave boundaries with their respective labels at the top of the plot. The shaded portion is not considered in this study as suggested in [34]. Table I details the frequencies used for each sub-band.

The total feature vector per trial, $\mathbf{X}$, is then constructed in the following way:

$$\mathbf{X} = [\mathbf{X}_1^T \quad \mathbf{X}_2^T \quad \cdots \quad \mathbf{X}_{32}^T]^T \tag{9}$$

where $\mathbf{X}_\ell \in \mathbb{R}^{10}$, $\ell = 1, 2, ..., 32$, is the feature vector for each of the 32 EEG channels, containing the average power spectrum value across the 10 sub-bands defined above.

The feature selection model we implement is the Neighborhood Component Analysis (NCA) in MATLAB, implemented through the function *fscnca*, with the solver set as stochastic gradient descent, implemented using the option *sgd*. To select the final set of features used in classification, $\mathbf{F}$, we use the NCA model feature weights, $\mathbf{W}$, and a relative threshold factor, r. Features are selected based on the following criteria:

$$\mathbf{W} = [W_1^{(1)} \quad W_1^{(2)} \quad \cdots \quad W_1^{(h)} \quad \cdots \quad W_q^{(h)}]^T \tag{10}$$

$$\mathbf{F} \subseteq \mathbf{X} \quad \forall \quad W_q^{(h)} > r \cdot \max([1 \quad \max(\mathbf{W})]) \tag{11}$$

where $W_q^{(h)}$, is the NCA model feature weight of channel q, $q = 1, 2, ..., C$, and sub-band h, $h = 1, 2, ..., S_B$, where $C = 32$ and $S_B = 10$, and $max(\cdot)$ returns the maximum scalar value of an array provided as input. In (11), the maximum value is first obtained from $\mathbf{W}$ and provided in a 2-element array, with the first element being 1. From here, we obtain the maximum value of this 2-element array and multiply it by r. $\mathbf{F}$ will contain a subset of the features $\mathbf{X}$ in (9) for any respective weights (i.e., $W_q^{(h)} \mapsto X_q^{(h)}$)) that are greater than the relative threshold calculated in (11). For this study, r is taken to be 0.02.

TABLE I: Sub-band Frequency Ranges

Waveband	Low Frequency (Hz)	High Frequency (Hz)
δ	0.5 – 1.75	1.75 – 4
θ	4 – 5.75	5.75 – 7.5
α	8 – 10.5	10.5 – 13
β	14 – 20	20 – 26
γ	30 – 37.5	37.5 – 45

For feature classification, we use the k-Nearest Neighbors (k-NN) classifier [35] using the MATLAB *fitcknn* function implementation. Initial analysis showed that a k-value of 1 would yield the best results. We stratified the training data and completed a 5-fold cross-validation, standardizing the predictors for the k-NN model implementation. To compare classification performance, we used Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM) classifiers using the MATLAB *fitcdiscr* and *fitcsvm* functions, respectively.

III. Results

Given the subjectivity of perceived trust, the results will be shown initially per subject. Additionally, since the subjects were exposed to three different swarming behaviors, we compare the results of a classifier trained initially on three classes (behaviors). However, we then group the behaviors into low and high levels of trust and present the results for binary classification as well. We will then consolidate the results and show what generalizations can be made about the human trustworthiness of robot swarms using EEG features.

A. Low Trust vs High Trust: Brain Topographies

Trust is an important relationship in human-swarm interaction. The human should be able to trust that the swarm is addressing the task at hand. As such, we set out to determine if we could perceive different levels of trust at the EEG level when observing different swarming behaviors that are related to the ability to execute the task at hand. We categorized *flock* and *surround* as high-trust since these swarms always reach the goal, and the *explore* behavior as low-trust as this swarm never reaches the goal, based on the criteria noted in Section II. The categorization is consistent with the post-interview assessment, where we asked each subject to describe their experiences with the simulations. All six subjects mentioned trusting the red (*surround*) and blue (*flock*) swarm to reach the goal on the screen while distrusting the green (*explore*) swarm.

In order to compare EEG recordings between those two groups of high and low trust, we applied a standard t-test to our low-trust (*explore*) and high-trust (*flock* and *surround*) EEG features. The goal is to test significant differences between low- and high-trust levels. Figure 6 shows the brain topographies of the significance across sub-bands of all six subjects. At each sub-band, and for each channel, a standard t-test was applied to compare the power spectrum density of low-trust and high-trust, where low trust is associated with the *explore* behavior and high trust to both *flock* and *surround* behaviors. An average of statistical significance, shown in color, is calculated across sub-bands with a value of 1 showing that each frequency step in the sub-band showed significance and a value of 0 showing that no frequency in the sub-band showed significance. Significance is defined based on the standard p-value, $p < 0.05$. Fig. 6 shows that each subject experiences the difference in low and high trust scenarios

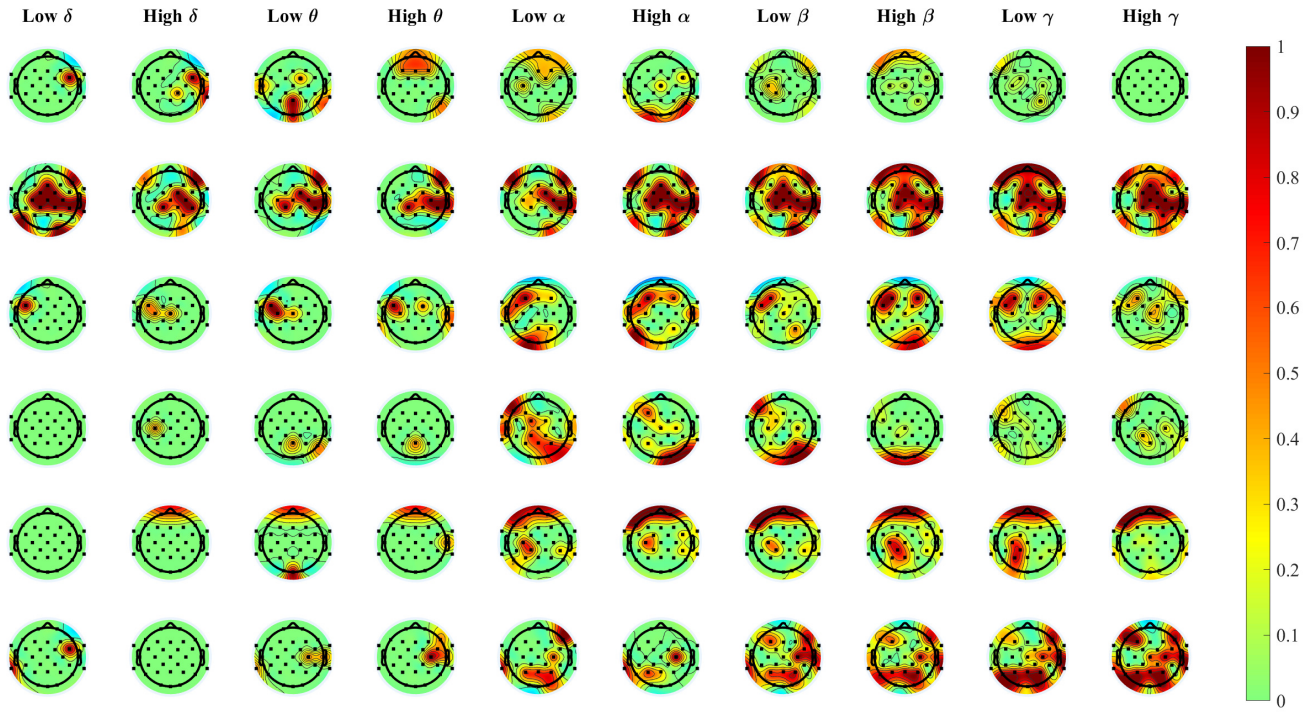

Fig. 6: Brain topography of all six subjects (rows) across all 10 sub-bands (columns), as defined in Table I. At each sub-band, a standard t-test was applied to compare the power spectrum density of low-trust and high-trust, where low trust is associated with the *explore* behavior and high trust to both *flock* and *surround* behaviors. An average of significance, shown in color, is calculated across sub-bands with a value of 1 showing that each frequency step in the sub-band showed significance and a value of 0 showing that no frequency in the sub-band showed significance. Dots on the brain map represent electrodes or EEG channels. Significance is defined based on the standard p-value, $p < 0.05$.

differently. For example, subject 2 (second row) experienced a higher perceived trust difference, whereas subjects 1 and 4 demonstrated fewer differences in significance. However, most of the differences observed occurred in the alpha, beta, and gamma bands across all six subjects, except for subject 2 where there was a more significant difference in low delta compared to low alpha. Moreover, if we consider the spatial distribution of the features that showed statistically significant differences between the two levels of trust, we see most of the differences in the frontal and occipital regions across all subjects.

B. Feature Classification

While noting differences in brain response when exposed to different swarming behaviors is essential, we need to establish if we can use this information to identify EEG features of trust that would allow us to anticipate user distrust of a swarm and update a swarm's control, for example. We trained a model described in Section II using low and high trust as the classifiers, using leave-one-out five-fold cross-validation for each behavior. Figure 7 shows the confusion matrix of the aggregate classification accuracy across all subjects. We see a 98.4% and 100.0% prediction accuracy for high- and low-trust classes, respectively. Only one low-trust trial was misclassified as high-trust (subject 1), while all high-trust classes were correctly predicted, across all subjects. Additionally, we compared the results of the k-NN classifier with LDA and SVM classifiers. Table II shows the classification results of all three classifiers. We see that the k-NN classifier outperforms both LDA and SVM classifiers in predicting both high-trust and low-trust classes. While classification results for high-trust are comparable (i.e., within 2% of each other), we see that k-NN outperforms both LDA and SVM by approximately 17% - 10% for low-trust classification, respectively.

TABLE II: k-NN, LDA, and SVM Feature Classification Results. Combined classification accuracy of all six subjects based on three different classifiers.

Classifier	k-NN	LDA	SVM
High-Trust	98.4 %	98.2 %	96.6 %
Low-Trust	100.0 %	82.9 %	90.3 %

Although we initially grouped the three behaviors into two levels of trust and presented the binary classifier above, we analyzed the performance of a classifier to the three behaviors *explore*, *flock*, and *surround* too. Similarly to the binary classifier, we trained and tested our classification model based on the three swarm behaviors. As seen in Fig. 8, we achieved a 96.8%, 84.4%, and 96.2% classification prediction accuracy for *explore*, *flock*, and *surround* behaviors, respectively. The

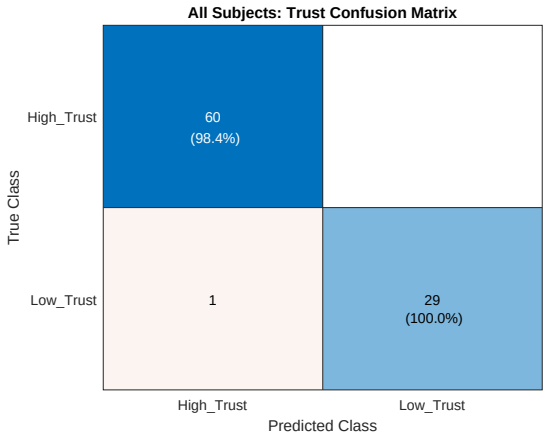

Fig. 7: Combined confusion matrix of all six subjects using the low- and high-trust classes. The model was trained and tested for each subject with the classification results added together. Out of 90 total observations, the low-trust and high-trust conditions were correctly predicted 100.0% and 98.4% of the time, respectively.

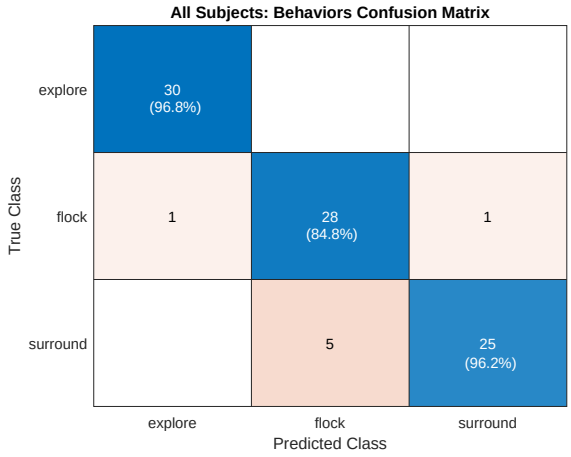

Fig. 8: Combined confusion matrix of all six subjects using the three behaviors as separate classes. The model was trained and tested for each subject with the classification results added together. Out of 90 total observations, the *explore*, *flock*, and *surround* behaviors were correctly predicted 96.8%, 84.8%, and 96.2% of the time, respectively.

largest misclassification occurs with *surround*, where it was incorrectly classified as *flock* in some cases. Specifically, two instances of *flock* get misclassified (one to *explore*, one to *flock*), while *surround* gets misclassified as *flock* five times. Based on subject responses and that these behaviors always reach the goal, this misclassification was expected.

C. Similarity of Features Across Subjects

The above classifiers were trained to subject-specific data and features selected using the feature selection method presented in Section II. However, we wanted to analyze commonalities in features selected among subjects. For this reason, we analyzed which features were selected in each of the six subjects and looked for commonalities. Figure 9 shows the feature heatmap we use to determine which features are shared among all subjects. The rows represent each of the 10 features per channel, and the columns represent the 32 EEG channels. The number on each element represents the number of subjects for which the specific feature was selected by the algorithm noted in Section II. As seen in Fig. 9, there is no single feature shared among all six subjects, but there are six features that at least four subjects share. These are not the same four subjects for all six features. For example, Table III shows how only subjects 2 and 6 contain five of the six common features, while subjects 4 and 5 only contain three. As can be seen, each of these common features shows up at most in four subjects.

TABLE III: Common Features By Subject. Features are noted as $X_{a,b}$, where a, b correspond to their respective row and column placement in the feature heatmap in Fig. 9.

Subject	Subset of Features Selected per Subject
1	$X_{3,6}$, $X_{9,22}$, $X_{1,26}$, $X_{7,26}$
2	$X_{3,6}$, $X_{8,18}$, $X_{9,21}$, $X_{9,22}$, $X_{7,26}$
3	$X_{3,6}$, $X_{8,18}$, $X_{9,22}$, $X_{1,26}$
4	$X_{8,18}$, $X_{9,21}$, $X_{9,22}$
5	$X_{9,21}$, $X_{1,26}$, $X_{7,26}$
6	$X_{3,6}$, $X_{8,18}$, $X_{9,21}$, $X_{1,26}$, $X_{7,26}$

Common Features
$X_{3,6}$, $X_{8,18}$, $X_{9,21}$, $X_{9,22}$, $X_{1,26}$, $X_{7,26}$

Based on these results, we trained a classifier across all subjects, using only the six common features as inputs, to predict low or high trust. Figure 10 shows the results of this classification in cross-validation. It can be seen that low trust gets misclassified as high-trust four times, while high trust gets misclassified as low-trust seven times. However, the overall classification rates are very high for high trust, 93%, and satisfactory for low trust, 78.8%, which shows that only the six most common features across subjects can be used to predict trust levels with high accuracy across all subjects.

IV. Discussion

The results of this study suggest that EEG correlates of swarm trust exist and are distinguishable in machine learning feature classification, given the test conditions and swarm behaviors exhibited. This has the potential to impact the future of human-swarm teaming. The rest of the discussion explores the specific brain regions of trust found in this study, lists any shortcomings, and considers the future applications of this study's findings.

A. Brain Regions Related to Trust

Based on this study's results, we see that brain areas that were activated differently between the two levels of trust tested, as depicted in Fig. 6, are consistent with brain regions responsible for trust identified in previous studies [17]. Specifically, we see that there are higher significance differences between low and high trust scenarios in the left, central, and right frontal regions as well as the occipital region. This strengthens our claim that the observed brain activation

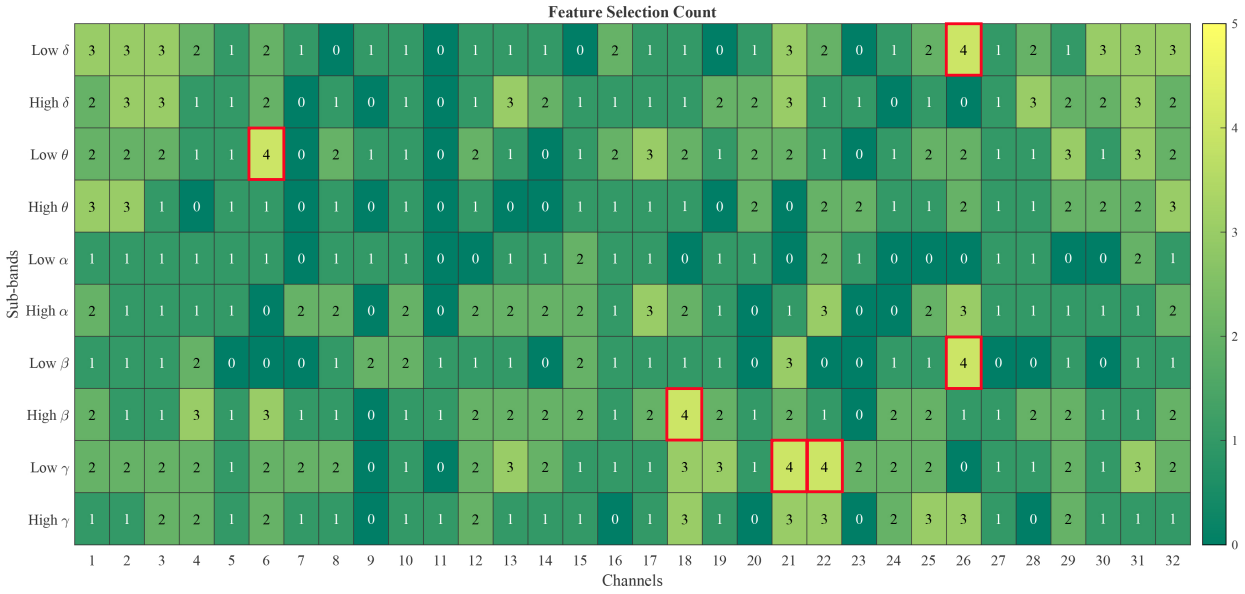

Fig. 9: Commonality of selected features across all six subjects represented as a heatmap. The features are described by sub-band (rows) and channel (columns). Each feature or cell a subject has will be counted as one. The maximum value a cell can have is six. The common features among 4 subjects are highlighted by a red box around them.

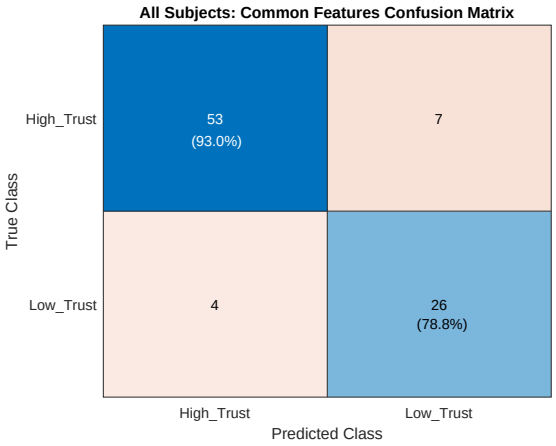

Fig. 10: Combined confusion matrix of all six subjects using the trust classes. The model was trained with the six most common features, see Table III, found in at most four subjects. The classification results are added together. Out of 90 total observations, the low-trust and high-trust conditions were correctly predicted 78.8% and 93.0% of the time, respectively.

differences are indeed due to different levels of trust, and they are not evoked due to specific parameters of the experiment, e.g., the motion of the swarm. When looking further for similarities across subjects, Fig. 11 shows the brain regions of the most common features discussed in Table III. We see that EEG channels FC5 and O2 are consistent with the previous findings in [17]. Additionally, we see that EEG channels T8 and TP10 are consistent with the work in [36], which finds that the temporal region in the gamma waves indicates levels of mistrust. In our case, only TP10 and CP6 are in the γ band. The results related to the most common features are promising, but we do not claim to be definitive. Additional work with many more subjects is needed to have definitive evidence that general features can be used with the majority of subjects with consistent and positive results.

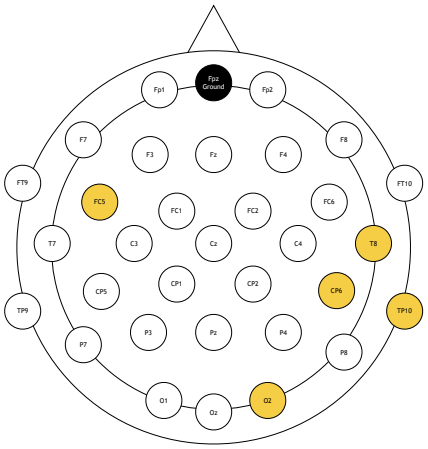

Fig. 11: The channel location of the most common features as seen in Fig. 9 and Table III. It is shown that left-central, occipital, centro-parietal, and temporal regions play a role in accurately determining trust levels in this study.

B. EEG Features Dimensionality Reduction

This study focused on a binary classification of trust (low and high) using three swarming behaviors that affect the trustworthiness of the swarm. Focusing on the prediction accuracy using a k-NN classifier, we obtain a high prediction accuracy across all subjects, as shown in Fig. 7. The model prediction accuracy is based on 72 ± 13 features out of 320 features per subject. We also found that we can drastically reduce the dimensionality by focusing on the most common features, resulting in 4 ± 1 features per subject, as seen in Table III,

and achieving great prediction accuracy, as seen in Fig. 10. This result is very promising because it can lead to drastically reducing the setup time for EEG electrode placement, and the computational cost, by focusing on the most common features and areas, while still achieving very accurate prediction results. For example, for one subject that used 64 of 320 features in classification, we could decrease the number of channels needed to be attached to the subject by about 1/3 (i.e., those 64 features came from only 20 channels). As such, we reduce the data being processed, thus reducing the total number of features being computed as well as improve on the physical setup time of attaching electrodes to the subject. While more work is needed to establish definitive general features across subjects as described earlier, we can see that the savings are even more drastic as we go from 20 channels to 6.

C. Limitations and Future Applications

Although an EEG interface does not seem to be practical for some applications, e.g., for a human operator of a swarm in a search and rescue scenario, we believe that the technological advances in EEG recording interfaces and systems will render this setup feasible and economical in the very near future. Moreover, our work shows that a minimum set of channels can lead to a high classification accuracy of trust, which can further help with the feasibility and simplicity of the proposed setup in future applications. As noted in the results section of this paper, we can achieve a high classification of trust given our experimental setup. We believe this to be the case given that trust is a mental construct that each individual can experience differently, compared to other work in Brain-Computer Interfaces (BCIs) where a subject actuates a robot arm or manipulates a swarm into a preset configuration [37], [38]. Future work could explore the reproducibility of such high classification rates found here through testing the same subjects on different days, for example. However, as with other BCIs, retraining the classifier has to be done every time it is intended to be used. Thus, testing the same subjects on different days does not mean we use the same trained model from day one, rather, we need to retrain the model every day the subject is tested.

The proposed work introduces EEG-based feature classification related to human trust in robotic swarms for the first time, which has direct implications for effective human-machine teaming with applications to many fields such as exploration, search and rescue operations, surveillance, environmental monitoring, and defense. In those applications, quantifying levels of human trust in the deployed swarm is of utmost importance because it can lead to swarm controllers that adapt their output based on the human's perceived trust level. In these scenarios where the swarm can exhibit different swarming behaviors to adjust to the task (e.g., overcoming obstacles, changing configurations, etc), it is important that the subject can discern between a swarm's disregard for a command instead of the swarm adapting to its current environment. As such, future work could focus on changing a swarm's behavior within the trial either working towards or against the task at hand. Moreover, applications in which the operator's hands are constrained (e.g., active conflict, search-and-rescue, on-orbit maintenance, manufacturing) would benefit from this type of system. Operators could continue their work without manually operating a swarm's current target, which might increase productivity. This type of interaction would require an operator's attention to shift depending on the task. The work discussed here depends on the subject focusing on and thinking about the swarm's trustworthiness throughout the entirety of the trials. Future work could focus on determining if trust levels are discernable when subjects are preoccupied with additional tasks. Additionally, providing transparency in human-machine teaming is an effective way to establish comprehension of a given task of a collective behavior [39], [40]. Leveraging the findings discussed in this paper, we can design an interface that provides a swarm's current understanding of a task and an operator's current trust levels, in order to provide an operator insight into the swarm's current intentions. Fig. 12 provides an example of how we could use this for search-and-rescue operations. The human operator can safely and remotely observe the swarm and provide feedback through EEG signals if the autonomous agents are putting survivors in danger.

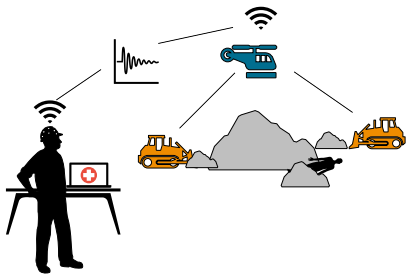

Fig. 12: Example of a human operator in the loop overseeing rubble removal. The operator is able to observe the process and indirectly provide feedback with brain signals.

V. Conclusion

Trust is a barrier to artificial intelligence technology adoption [2], [3]. This paper establishes neural correlates of human trust when observing a robot swarm complete a task. In this study, EEG data were used to find differences in trust using three swarming behaviors (*explore*, *flock*, and *surround*) to reach a target. The results of this paper show that we can identify a low trust and high trust scenario to an increased level of accuracy for each subject with approximately only 22% of the available EEG data. Additionally, we show that the dimensionality of the necessary EEG features can be decreased even further to only six features, which correspond to approximately 2% of the available data. Using only this small set of features, the human trust levels in the swarm can still be identified with high accuracy. To the best of our knowledge, this is the first time that EEG measurements are used to find trust features when dealing with a robot swarm exhibiting different collective behaviors. This work has direct implications for effective human-machine teaming with applications to many fields such as exploration, search and

rescue operations, surveillance, environmental monitoring, and defense. In those applications, quantifying levels of human trust in the deployed swarm is of utmost importance because it can lead to swarm controllers that adapt their output based on the human's perceived trust level.

REFERENCES

[1] M. Alessandro, B. Cardinale Lagomarsino, C. Scartascini, J. Streb, and J. Torrealday, "Transparency and trust in government. evidence from a survey experiment," *World Development*, vol. 138, p. 105223, 2021.

[2] L. Christoforakos, A. Gallucci, T. Surmava-Große, D. Ullrich, and S. Diefenbach, "Can robots earn our trust the same way humans do? a systematic exploration of competence, warmth, and anthropomorphism as determinants of trust development in hri," *Frontiers in Robotics and AI*, vol. 8, 2021.

[3] K. Kaur and G. Rampersad, "Trust in driverless cars: Investigating key factors influencing the adoption of driverless cars," *Journal of Engineering and Technology Management*, vol. 48, pp. 87–96, 2018.

[4] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, 2004.

[5] C. W. Reynolds, "Flocks, herds and schools: A distributed behavioral model," *ACM SIGGRAPH Computer Graphics*, vol. 21, no. 4, pp. 25–34, 1987.

[6] G. K. Karavas and P. Artemiadis, "On the effect of swarm collective behavior on human perception: Towards brain-swarm interfaces," in *2015 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, 2015, pp. 172–177.

[7] G. K. Karavas, D. T. Larsson, and P. Artemiadis, "A hybrid bmi for control of robotic swarms: Preliminary results," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5065–5075.

[8] P. Walker, S. Amirpour Amraii, M. Lewis, N. Chakraborty, and K. Sycara, "Control of swarms with multiple leader agents," in *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, Conference Proceedings.

[9] J. Nagi, A. Giusti, L. M. Gambardella, and G. A. Di Caro, "Human-swarm interaction using spatial gestures," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 3834–3841.

[10] M. Chen, P. Zhang, X. Chen, Y. Zhou, F. Li, and G. Du, "A human-swarm interaction method based on augmented reality," in *2018 WRC Symposium on Advanced Robotics and Automation (WRC SARA)*. IEEE, 2018, pp. 108–114.

[11] S. Choo and C. S. Nam, "Detecting human trust calibration in automation: A convolutional neural network approach," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 4, pp. 774–783, 2022.

[12] W.-L. Hu, K. Akash, T. Reid, and N. Jain, "Computational modeling of the dynamics of human trust during human–machine interactions," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 6, pp. 485–497, 2019.

[13] F. Ekman, M. Johansson, and J. Sochor, "Creating appropriate trust in automated vehicle systems: A framework for hmi design," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 1, pp. 95–101, 2018.

[14] C. Nam, P. Walker, H. Li, M. Lewis, and K. Sycara, "Models of trust in human control of swarms with varied levels of autonomy," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 3, pp. 194–204, 2020.

[15] B. W. Haas, A. Ishak, I. W. Anderson, and M. M. Filkowski, "The tendency to trust is reflected in human brain structure," *NeuroImage*, vol. 107, pp. 175–181, 2015.

[16] K. F. Firoz, Y. Seong, and S. Oh, "A neurological approach to classify trust through eeg signals using machine learning techniques," in *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)*, 2022, pp. 1–6.

[17] M. Wang, A. Hussein, R. F. Rojas, K. Shafi, and H. A. Abbass, "Eeg-based neural correlates of trust in human-autonomy interaction," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, Conference Proceedings.

[18] W.-L. Hu, K. Akash, N. Jain, and T. Reid, "Real-time sensing of trust in human-machine interactions," *IFAC-PapersOnLine*, vol. 49, no. 32, pp. 48–53, 2016.

[19] M. Seet, J. Harvy, R. Bose, A. Dragomir, A. Bezerianos, and N. Thakor, "Differential impact of autonomous vehicle malfunctions on human trust," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 548–557, 2022.

[20] K. Gupta, R. Hajika, Y. S. Pai, A. Duenser, M. Lochner, and M. Billinghurst, "Measuring human trust in a virtual assistant using physiological sensing in virtual reality," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2020, pp. 756–765.

[21] G. H. Klem, "The ten-twenty electrode system of the international federation. the international federation of clinical neurophysiology," *Electroencephalogr. Clin. Neurophysiol. Suppl.*, vol. 52, pp. 3–6, 1999.

[22] A. Singh and P. Artemiadis, "Automatic identification of the leader in a swarm using an optimized clustering and probabilistic approach," in *2021 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, 2021, pp. 1–6.

[23] S. Nagavalli, S.-Y. Chien, M. Lewis, N. Chakraborty, and K. Sycara, "Bounds of neglect benevolence in input timing for human interaction with robotic swarms," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 197–204.

[24] C. Kothe, "Lab streaming layer," 2013. [Online]. Available: https://labstreaminglayer.readthedocs.io/

[25] F. Sarembaud, https://github.com/cromod/Collision, 2016.

[26] P. Walker, M. Lewis, and K. Sycara, "Characterizing human perception of emergent swarm behaviors," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016, pp. 002 436–002 441.

[27] R. Sorbello, S. Tramonte, M. E. Giardina, V. La Bella, R. Spataro, B. Allison, C. Guger, and A. Chella, "A human–humanoid interaction through the use of bci for locked-in als patients using neuro-biological feedback fusion," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 2, pp. 487–497, 2018.

[28] D. Wijayasekara and M. Manic, "Human machine interaction via brain activity monitoring," in *2013 6th International Conference on Human System Interactions (HSI)*, 2013, pp. 103–109.

[29] K. Nakamura and K. Natsume, "Detection of error-related potentials during the robot navigation task by humans," in *2020 International Conference on Computational Intelligence (ICCI)*, 2020, pp. 153–158.

[30] K. Schaaff and T. Schultz, "Towards an eeg-based emotion recognizer for humanoid robots," in *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, 2009, pp. 792–796.

[31] S. Bozinovski and A. Bozinovski, "Mental states, eeg manifestations, and mentally emulated digital circuits for brain-robot interaction," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 1, pp. 39–51, 2015.

[32] D. Arnaud and M. Scott, "Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, pp. 9–21, 2004.

[33] P. He, G. Wilson, and C. Russell, "Removal of ocular artifacts from electro-encephalogram by adaptive filtering," *Medical & Biological Engineering & Computing*, vol. 42, no. 3, pp. 407–412, 2004.

[34] S. Sanei and J. A. Chambers, *EEG signal processing and machine learning*. John Wiley & Sons, 2021.

[35] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[36] S. Oh, Y. Seong, S. Yi, and S. Park, "Investigation of human trust by identifying stimulated brain regions using electroencephalogram," *ICT Express*, vol. 8, no. 3, pp. 363–370, 2022.

[37] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Adaptive multi-degree of freedom brain computer interface using online feedback: Towards novel methods and metrics of mutual adaptation between humans and machines for bci," *PloS one*, vol. 14, no. 3, p. e0212620, 2019.

[38] H. Wang, X. Dong, Z. Chen, and B. E. Shi, "Hybrid gaze/eeg brain computer interface for robot arm control on a pick and place task," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 1476–1479.

[39] K. A. Roundtree, J. R. Cody, J. Leaf, H. O. Demirel, and J. A. Adams, "Transparency's influence on human-collective interactions," *ACM Transactions on Human-Robot Interaction*, vol. 11, no. 2, pp. 1–48, 2022.

[40] A. J. Hepworth, D. P. Baxter, A. Hussein, K. J. Yaxley, E. Debie, and H. A. Abbass, "Human-swarm-teaming transparency and trust architecture," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, pp. 1281–1295, 2021.