ARL-TR-10229 ● Nov 2025

Multimodal Predictors of Team Trust and Task Cohesion Measured During an Experiment Examining Dynamic Task Allocation Technologies in a Next Generation Combat Vehicle Simulator

by David Chhan, Murat Kucukosmanoglu, Catherine Neubauer, and Andrea Krausman

NOTICES

Disclaimers

ARL-TR-10229 ● NOV 2025

Multimodal Predictors of Team Trust and Task Cohesion Measured During an Experiment Examining Dynamic Task Allocation Technologies in a Next Generation Combat Vehicle Simulator

David Chhan, Catherine Neubauer, and Andrea Krausman
DEVCOM Army Research Laboratory

Murat Kucukosmanoglu
D-Prime LLC

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED START DATE	END DATE
November 2025	Technical Report	10/1/2024	9/30/2025

4. TITLE AND SUBTITLE
Multimodal Predictors of Team Trust and Task Cohesion Measured During an Experiment Examining Dynamic Task Allocation Technologies in a Next Generation Combat Vehicle Simulator

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
David Chhan, Murat Kucukosmanoglu, Catherine Neubauer, and Andrea Krausman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
DEVCOM Army Research Laboratory
ATTN: FCDD-RLA-FE
Aberdeen Proving Ground, MD 21005

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-10229

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
ORCIDs: David Chhan, 0000-0003-2470-8663; Catherine Neubauer, 0000-0002-6686-3576; Andrea Krausman, 0000-0003-1955-8867

14. ABSTRACT
This report summarizes recent analyses aimed at identifying indicators of team trust and cohesion by examining relationships between subjective survey data and physiological and behavioral measures in support of the Scalable Cross-Echelon Command and Control Program. These analyses are part of a broader effort to develop metrics and predictive models for continuous assessment of human–autonomy teams and understand how to apply these models within the context of future Soldier and AI-enabled command and control teams. Data for the analyses were collected during a simulation involving 30 Soldiers conducting 26 operational missions in a vehicle simulator. The goal of the study was to test and validate the dynamic re-tasking and task allocation technology developed to assist Soldiers in effectively managing resources and improving situation awareness. Key findings revealed significant correlations between individual self-reported trust and cohesion responses (trust-in-team, trust-in-technology, and task cohesion) and variables, such as communication sentiment, cardiac activities (heart rate and heart rate variability), and oculomotor behaviors (pupil and gaze measures). We found eye-based metrics aggregated at the team level provided the most consistent and strongest correlation. Using the features derived from physiological and behavioral measures, machine learning models were developed to predict individual team trust, trust-in-technology, and task cohesion. The models performed moderately well in predicting the self-reported trust and cohesion responses with good correlation between the predicted and true values where the difference between the two values falls within a 15% range. Although based on a relatively small sample size, these results can provide useful insights and inform the design of systems to enhance the performance of both human–human and human–autonomy teams in future manned–unmanned teaming operations.

15. SUBJECT TERMS
Humans in Complex Systems, human–autonomy teaming, team trust and cohesion, multimodal predictive modeling, physiological and behavioral measures, continuous assessment

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	C. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UU	40

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
David Chhan	(520) 691-4778

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. Introduction

The rapidly advancing capabilities of autonomy, robotics, and artificial intelligence (AI)—including large language models, generative AI, and machine learning—present opportunities to reduce Soldier burden, accelerate data processing, increase situational awareness, and improve combat effectiveness. The U.S. Army is focused on effectively integrating autonomy and AI as collaborative team members to fully realize these benefits. Achieving this requires creating synergistic human–autonomy teams, necessitating research into the team dynamics that govern their performance—particularly, the critical constructs of trust and cohesion.

- Team Trust: Defined as team members' positive expectations of mutual competence and motivation, and a shared acceptance of vulnerability.[1] Trust is a dynamic state that fundamentally influences team processes, such as information sharing and mutual support.[2,3] In the context of human–autonomy teaming, proper trust calibration is essential to mitigate issues of overreliance or underutilization of autonomous systems.[4,5]
- Team Cohesion: Described as the shared bond that unifies a team in pursuit of its objectives.[6–8] Cohesion comprises both task cohesion (motivation to achieve group goals) and social cohesion (motivation to maintain interpersonal relationships) and is a primary driver of team motivation and performance.[9]

To effectively assess these constructs in heterogeneous teams consisting of human and autonomous/robotics systems, we have developed and tested a multimodal assessment approach[10] and software toolkit,[11,12] which integrates subjective self-reports with continuous objective data streams, including high-dimensional mission/system data and physiological/behavioral data from wearable sensors (e.g., heart rate monitors, eye trackers). This fusion of data provides a comprehensive time-synchronized understanding of team states.

This report documents the data analyses performed to further validate our multimodal approach for the continuous assessment of team trust and cohesion. We analyze physiological and behavioral measures identified in the literature as correlates of these states. The goal is to further develop and validate predictive models of trust and cohesion, enabling real-time measurement and targeted interventions to ensure optimal team calibration in operational settings and to understand how these assessment methods extend to other operational contexts in which Soldiers interact with advanced technologies, such as Next Generation Combat Vehicle (NGCV) command and control teams in support of the Scalable Cross-Echelon Command and Control (SC3) Program.

2. The Study

The analyses described in this report are based on data collected during the experiment "Technologies for Automated Dynamic Resource Allocation in the Next Generation Combat Vehicle."[13] The protocol for this experiment was approved by the U.S. Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory (ARL) Institutional Review Board. The primary objective of the experiment was to test and validate the dynamic re-tasking and task allocation capability; namely, the modular-trigger response system (MTRS) that enables Soldiers to customize the dynamic tasking of crew members and the deployment of autonomous resources according to team state and mission context, set up alerts, and receive prompts for role switching. One of the autonomous resources is the computer-vision–based aided target recognition system called the Automatic Detection System (ADS). ADS provides enhanced situation awareness capability through target detection and identification. To examine the influence of the MTRS and usefulness of the ADS system, a simulation exercise was held. Platoons of Soldiers (i.e., the friendly forces or BLUFOR) used vehicle simulators to conduct missions against an opposing force (OPFOR) in a simulated environment. After training and familiarization with the systems, the Soldiers conducted 13 missions. The missions occurred at different locations in the virtual environment and included various combinations of common armor and cavalry missions, such as reconnaissance, movement to contact, attack, and defense. The influence of technologies on performance of the platoons is described in Gremillion et al.[13] In addition to collecting performance-related data from the simulation system, behavioral, physiological, and subjective data were collected from Soldier participants before, during, or after each experimental session.

In the experiment, the BLUFOR comprised Soldiers from the 1st Cavalry Division, 1st Armored Brigade Combat Team. A total of 30 Soldiers with experience in armor and cavalry operations participated. They participated as two separate groups of 15 Soldiers. Fourteen of the Soldiers filled the roles of crew members in the vehicle simulators. The 15th Soldier filled the role of Higher Control (HICON). This Soldier coordinated with the research staff, passed information to and from the crew, evaluated the crew's performance during the simulated missions, and led the after-action reviews. Data was not collected from the 15th Soldier. Each group participated during a different 2-week period. Members of the research team filled in the roles of the OPFOR.

The experiment took place in ARL's Information for Mixed Squads (INFORMS) laboratory at Aberdeen Proving Ground, Maryland. The INFORMS laboratory contains a platoon-level NGCV simulator. The simulator consists of two identical vehicle mock-ups. Each mock-up represents a manned combat vehicle (MCV) and contains seven crew stations (Figure 1). One crew station is for the vehicle commander. Two crew stations are for the gunners and drivers of the MCV, and the remaining crew stations are for the gunners and drivers of the two robotic combat vehicles (RCVs).

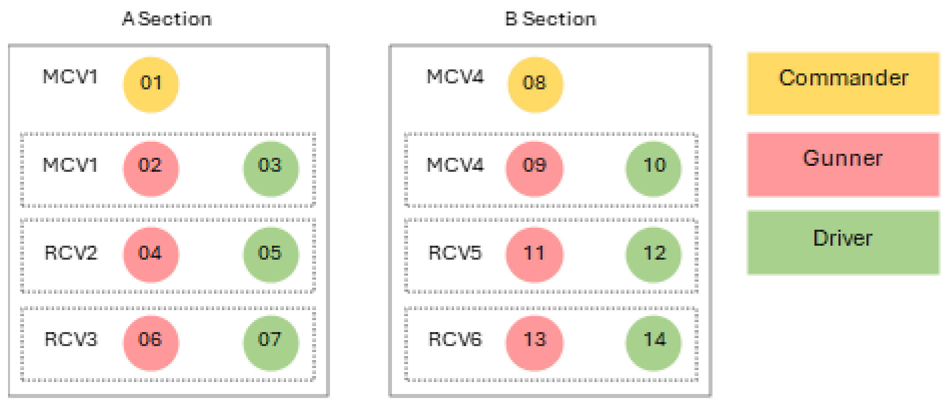

Figure 1. Schematic of the seating positions for the crew members in the NGCV simulators at the INFORMS laboratory.

Each group of subjects participated over a different 2-week period. The first 3 days of the first week were devoted to familiarization and training. Then, the focus shifted to the data collection during simulated missions. The missions were approximately 90 min long. Subjects typically completed two missions per day. The missions were broken into four phases. The first one or two phases typically involved movement and reconnaissance. The later phases involved engagements with the OPFOR as the subjects maneuvered to seize objectives.

3. Methods

The following equipment and instruments were used to collect behavioral, physiological, and subjective data during the experiment. These data were used to validate our multimodal approach for assessing team trust and cohesion. See Gremillion et al.[13] for details related to the larger study's hypotheses and results.

- Behavioral Measures
    - The JBL headsets, in addition to being used for communication among the crew stations, allowed researchers to capture verbal communications and use them in the analyses.
- Physiological Measures
    - Tobii Pro Glasses 3 (Tobii AB; Stockholm, Sweden) were used to measure pupil size, eye position, and gaze direction.

    - Zephyr BioHarness 3 (Medtronic Zephyr; Boulder, Colorado) was worn by subjects to capture electrocardiogram (ECG) and respiration data. Only the ECG data were used for analyses.
- Self-reported Measures
    - The Mid-Mission Probe Questionnaire has six questions pertaining to workload, team trust, and system trust. It was administered at the end of each phase in a mission. This report focuses on team trust (trust-in-team) and system trust (trust-in-ADS).
    - The Task Cohesion Questionnaire[14] was used to address the subjects' commitment to their tasks. This questionnaire contains five questions with ratings of 1–7. The overall task cohesion score is calculated by averaging all five items. This report focuses on task cohesion for the entire platoon.

Both the behavioral (e.g., speech) and physiological (e.g., ECG and eye-tracking) measures were recorded continuously during the mission. For self-reported metrics, the mid-mission probe questionnaire was administered at the end of each phase (i.e., four times in a mission), while the Cohesion Questionnaire was administered at the end of each mission.

3.1 Sentiment Feature Extraction

Communication-based audio data recorded during missions from each individual were processed and transcribed with the OpenAI Whisper speech recognition engine and segmented into 30-s non-overlapping windows. For each 30 s of transcribed communication in text data, sentiment features were extracted using a fine-tuned DistilBERT model developed in our prior work on sentiment analysis in military team communications.[15] The predicted probability scores of the three sentiment classes (positive, neutral, and negative) for each 30-s window served as the sentiment features.

The DistilBERT model was trained on military-specific communication data, capturing both emotional sentiment (e.g., frustration, satisfaction) and operational sentiment (e.g., mission success, tactical setbacks). For this study, predicted sentiment probabilities were obtained for each 30-s segment, forming a continuous time-series aligned with mission events. Table 1 provides an example output from the sentiment model for a mission communication segment.

Table 1. Example sentiment model output for a mission communication segment.

Communication segment	Sentiment class	Probability score
There, there, there. I see another one. And engaging. Two engaging at T90. Destroyed.	Negative (LABEL_0)	0.02
	Neutral (LABEL_1)	0.01
	Positive (LABEL_2)	0.97

3.2 ECG Feature Extraction

The ECG data were captured from each participant using a Zephyr BioHarness 3 (Medtronic Zephyr; Boulder, Colorado) chest strap device. The ECG data were sampled at 250 Hz and time-synchronized using the Apache Kafka protocol, a real-time data pipeline and streaming application. After time aligned with other data streams and the mission event timeline, the ECG signals were segmented into 30-s non-overlapping windows. Following the BioSPPy, the biosignal processing Python package, each 30-s ECG signal was band-pass filtered between 3 and 45 Hz to remove high-frequency noise and obtain distinctive *QRS complex* and *R* wave peaks of the ECG waveforms. Interbeat intervals were extracted to compute heart rate (HR), root-mean-square of successive differences (RMSSD), and standard deviation of the differences between adjacent *RR* intervals (SDSD). See Table 2. RMSSD and SDSD are time-domain heart rate variability (HRV) measures known to reflect the parasympathetic nervous system activity. Higher HRV indicates greater parasympathetic influence that is often associated with relaxation and recovery, whereas lower HRV suggests increases in stress and the fight-or-flight response.

Table 2. ECG-based features extracted per time window.

Feature	Description
HR	Heart rate
RMSSD	Root mean square of successive differences
SDSD	Standard deviation of the successive differences

3.3 Eye-Tracking Feature Extraction

Eye-tracking data were recorded using Tobii Pro Glasses at a sampling rate of 100 Hz, capturing time-stamped horizontal and vertical gaze coordinates along with pupil diameter measurements from both eyes. To enable high-resolution temporal analysis aligned with mission timelines, a custom Python-based pipeline was developed to extract behavioral and oculometric features from sliding windows of 10 s, with a 1-s step size.

For each window, a comprehensive set of features was computed across four categories: pupil dynamics, blink activity, saccades, and fixations. Blink events were defined as gaps in valid eye-tracking data lasting at least 300 ms, following conventions used in the PyGazeAnalyser toolkit[16] and related literature.[17] Pupil metrics included mean and standard deviation of diameter, as well as left–right inter-eye differences. Saccades were identified using inter-sample thresholds on velocity (>0.2 units/s) and acceleration (>30 units/s²) computed from gaze positions, which are normalized screen coordinates ranging from 0 to 1 (with [0,0] indicating the top-left, and [1,1] the bottom-right of the display). Valid saccades had durations between 20 and 300 ms. For each window, we computed saccade count, amplitude, and velocity statistics.

Fixation detection was based on spatial clustering of gaze points. Fixations were defined as sequences of points that remained within 0.05 normalized spatial radius (i.e., squared distance $\leq 0.05^2$) for a minimum of 100 ms, consistent with established parameters in eye movement research. The fixations are rarely less than 100 ms and often in the range of 200–400 ms. Extracted fixation features included count, mean, and variability of fixation duration, and spatial dispersion along the x and y axes. A summary of the extracted eye-tracking features and their descriptions is provided in Table 3.

Table 3. Eye-tracking features extracted per time window.

Feature category	Feature name	Description
Pupil dynamics	Pupil size Mean/Std	Average and variability of pupil diameter across both eyes
	Left–right diameter diff (Mean/Std)	Difference between left and right pupil diameters (mean and variability)
Blink activity	Blink count	Number of blinks, inferred from data gaps ≥300 ms
Saccades	Saccade count	Number of detected saccades per window
	Saccade distance (Mean/Std)	Amplitude of eye movements between fixations
	Saccade velocity (Mean/Std)	Speed of saccadic eye movements
Fixations	Fixation count	Number of fixations per window
	Fixation duration (Mean/Std)	Duration of gaze stabilizations on a single point
	Fixation coordinates (X/Y Mean/Std)	Spatial distribution of fixations (center and spread in x and y directions)

4. Results

In this section we report results for how different features from the different data modalities (e.g., communication, physiology, and behavioral) correlate with the self-reported trust and cohesion measures: team trust (trust-in-team), technology trust (trust-in-ADS), and task cohesion (platoon-task-cohesion). In addition, we separately developed machine learning models to predict team trust, trust-in-ADS, and task cohesion using those features listed above.

4.1 Trust-in-Team

Here, we report correlations between the phase-level self-reported trust-in-team and the multimodal behavioral and physiological features at both the individual and team level (i.e., the entire platoon).

4.1.1 Individual-Level Correlation Plots

4.1.1.1 Audio Communication Features: Predicted Scores for the Three Sentiment Classes (Negative, Neutral, and Positive) Generated by the Sentiment Model from Crew Audio Communications

Figure 2 presents scatter plots examining the relationship between predicted sentiment scores from the audio communication sentiment model (negative, neutral, and positive) and self-reported trust-in-team scores at the individual level. Across all three sentiment classes, the correlations with trust are negligible, with coefficients of $r = 0.00$ for negative sentiment, $r = 0.01$ for neutral sentiment, and $r = –0.03$ for positive sentiment. Negative and neutral sentiment scores show no discernible trend with trust, as data points are broadly distributed across trust levels, while positive sentiment displays a very slight negative relationship, indicating that higher positive sentiment scores are not consistently linked to higher trust ratings. Overall, these results suggest that sentiment derived from audio communications does not strongly predict trust-in-team ratings in this context.

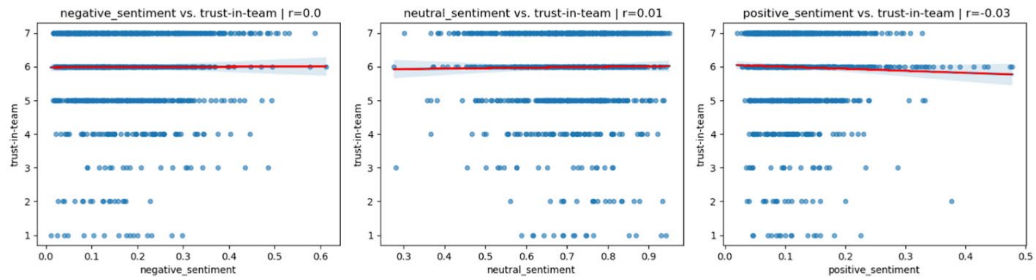

Figure 2. Correlations between sentiment scores features and trust-in-team ratings. Scatter plots display relationships among (left) negative sentiment, (center) neutral sentiment, and (right) positive sentiment with trust-in-team scores at the individual level. Red lines indicate linear regression fit, with correlation coefficients (*r*) and significance levels indicated in the plot titles.

4.1.1.2 ECG-based Features: HR, RMSSD, and SDSD

As shown in Figure 3, the scatter plot for HR shows a small but statistically significant positive correlation with trust-in-team (r = 0.13, p < 0.05), suggesting that higher HR values are modestly associated with higher trust ratings. In contrast, RMSSD, an indicator of parasympathetic activity and HRV, exhibits a small but significant negative correlation (r = –0.12, p < 0.05), indicating that higher variability is modestly linked to lower trust ratings in this context. Similarly, SDSD—which captures short-term variability between successive *RR* intervals—also shows a weak negative correlation (r = –0.08, p < 0.05), implying a slight tendency for greater short-term variability to be associated with lower trust scores. Overall, these relationships, though statistically significant, are small in magnitude, suggesting that while certain aspects of cardiac activity are related to trust, their predictive strength is limited.

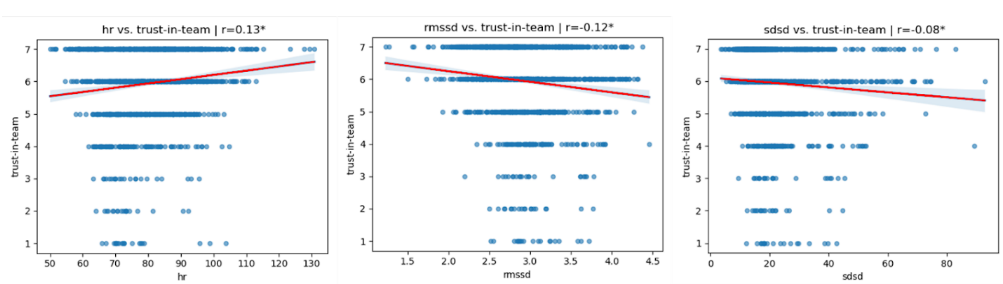

Figure 3. Correlations between ECG-derived features and trust-in-team ratings. Scatter plots display relationships among (left) HR, (center) RMSSD, and (right) SDSD with trust-in-team scores at the individual level. Red lines indicate linear regression fits, with correlation coefficients (*r*) and significance levels indicated in the plot titles.

4.1.1.3 Eye-based Features: Mean and Standard Deviation of Pupil Size, Saccade Count, Saccade Distribution, and Saccade Velocity

As shown in Figure 4, several eye-based metrics demonstrate small but statistically significant associations with trust-in-team ratings at the individual level. Pupil size mean ($r = –0.22$, $p < 0.05$), mean saccade velocity ($r = –0.26$, $p < 0.05$), and mean saccade distribution ($r = –0.35$, $p < 0.05$) all exhibit moderate negative correlations, indicating that larger pupil sizes, higher saccade velocities, and greater saccade distances tend to be associated with lower trust scores. Other features, including number of blinks ($r = –0.06$), fixation duration mean ($r = 0.00$), and saccade count ($r = –0.03$), show negligible relationships with trust. These findings suggest that certain oculomotor behaviors, particularly those linked to gaze dynamics and arousal, may be modestly related to perceived trust, although effect sizes remain too small to moderate.

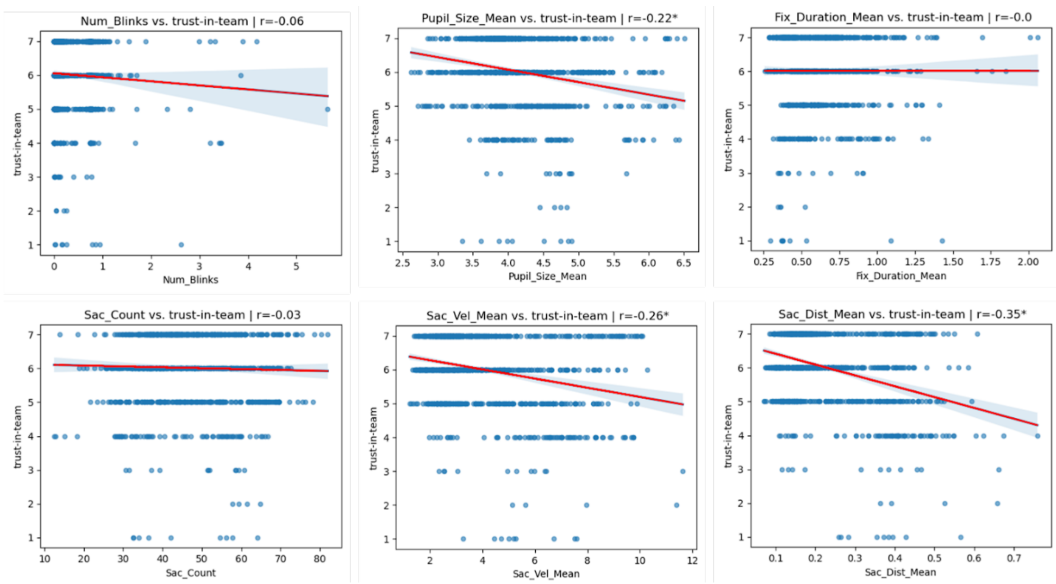

Figure 4. Individual level correlations between eye-based features and trust-in-team ratings. Scatter plots show relationships among (top row, left to right) number of blinks, mean pupil size, and mean fixation duration; and (bottom row, left to right) saccade count, mean saccade velocity, and mean saccade distribution with trust-in-team scores. Red lines indicate linear regression fits, shaded areas represent 95% confidence intervals, and correlation coefficients (*r*) with significance markers are displayed in each plot title.

4.1.2 Team-Level Correlation Plots

4.1.2.1 Team-Level Audio Communication Features

Figure 5 shows the team-level correlations between predicted sentiment scores from the audio communication sentiment model and trust-in-team ratings, where sentiment values are averaged across all crew members within each phase. Compared to the individual-level analyses, these aggregated results reveal stronger associations between communication sentiment and trust. Negative sentiment demonstrates a strong negative correlation (r = –0.55, $p < 0.05$), indicating that higher proportions of negative sentiment in team communications are linked to substantially lower trust ratings. In contrast, neutral sentiment shows a strong positive correlation (r = 0.48, $p < 0.05$), suggesting that higher levels of neutral sentiment are associated with higher trust. Positive sentiment displays a weaker positive correlation (r = 0.15), indicating only a modest relationship with trust ratings. Because outliers are suppressed in the aggregation process, these findings highlight the advantage of using the mean to mitigate the influence of extreme values and uniform response tendencies. Overall, the results suggest that at the team level sentiment composition in communications, particularly the balance between negative and neutral sentiment, is more predictive of trust than at the individual level.

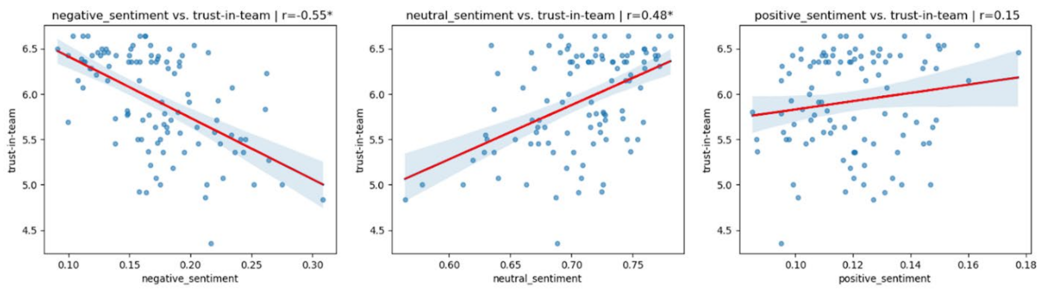

Figure 5. Team-level correlations between communication sentiment features and trust-in-team ratings. See Figure 2 for plot layout and description details.

4.1.2.2 Team-Level ECG-based Features

Figure 6 shows the team-level correlations between ECG-derived features and trust-in-team ratings, where physiological values were averaged across all crew members within each phase. HR displays a strong positive correlation with trust (r = 0.54, $p < 0.05$), suggesting that teams with higher average HRs tend to report higher trust levels. In contrast, RMSSD, an index of parasympathetic activity and HRV, shows a moderate negative correlation (r = –0.25, $p < 0.05$), indicating that teams with greater variability tend to have lower trust ratings. Similarly, SDSD—a short-term variability metric—is negatively correlated with trust (r = –0.23, $p < 0.05$). Because averaging suppresses outliers, these findings highlight the advantage of using team-level aggregation to reduce the influence of extreme physiological readings and response biases. Overall, the results suggest that at the team level, cardiac measures, particularly HR, are more strongly associated with trust than observed in individual-level analyses.

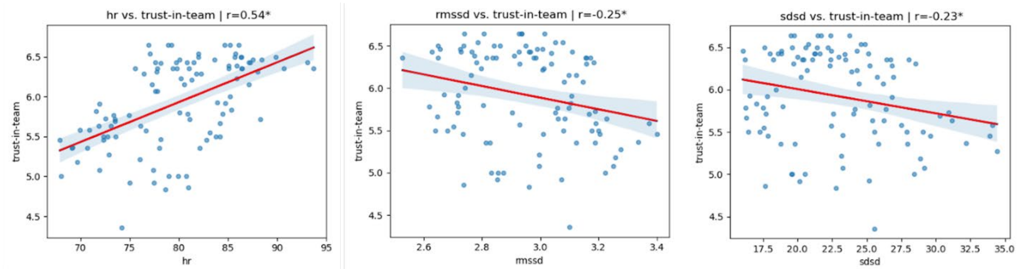

Figure 6. Team-level correlations between ECG-derived features and trust-in-team ratings. See Figure 3 for plot layout and description details.

4.1.2.3 Team-Level Eye-based Features

Figure 7 shows the team-level correlations between eye-based metrics and trust-in-team ratings, where feature values were averaged across all crew members within each phase. Several features exhibit strong and statistically significant correlations. Pupil size is positively associated with trust (r = 0.42, p < 0.05), indicating that teams with larger average pupil sizes tend to report higher trust. Conversely, pupil size standard deviation shows a strong negative correlation (r = –0.64, p < 0.05), suggesting that greater variability in pupil size is linked to lower trust ratings. Saccade count demonstrates a strong positive correlation (r = 0.52, p < 0.05), while mean saccade distance (r = –0.60, p < 0.05), mean saccade velocity (r = –0.58, p < 0.05), and saccade velocity standard deviation (r = –0.49, p < 0.05) all show strong negative relationships with trust. Because aggregation suppresses outlier effects, these findings underscore the benefit of using averaged values to better capture team-level patterns. Overall, the results indicate that gaze dynamics—especially those related to saccade behavior and pupil size variability—are strongly associated with trust at the team level.

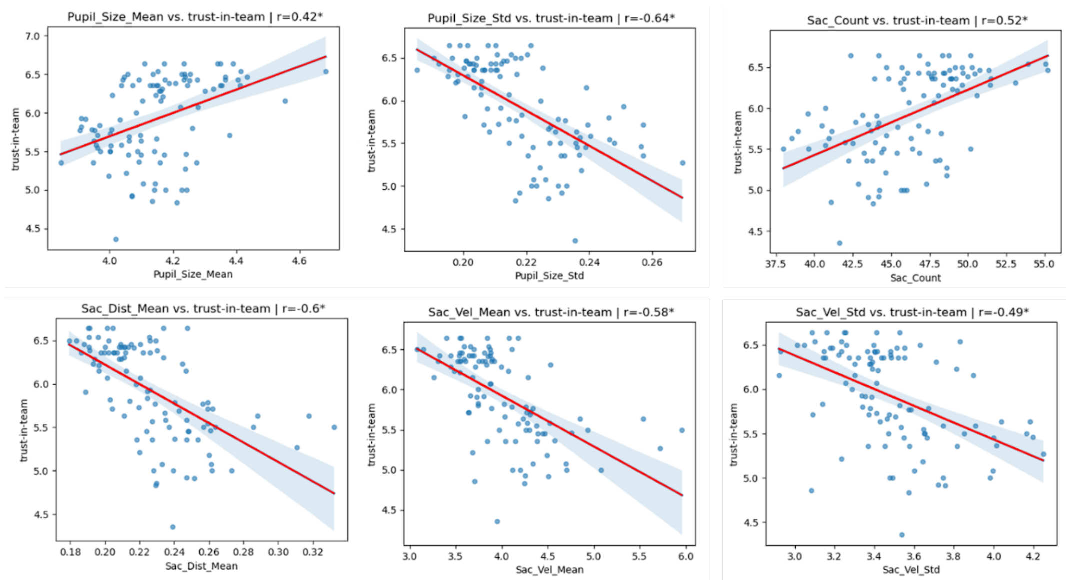

Figure 7. Team-level correlations between eye-based features and trust-in-team ratings. See Figure 4 for plot layout and description details.

4.2 Trust in the ADS

Here, we report correlations among the phase-level self-reported trust in ADS and the multimodal behavioral and physiological features at both the individual and team levels.

4.2.1 Individual-Level Correlation Plots

4.2.1.1 Audio Communication Features: Predicted Scores for the Three Sentiment Classes (Negative, Neutral, and Positive) Generated by the Sentiment Model from Crew Audio Communications

Figure 8 shows scatter plots of predicted sentiment scores for negative, neutral, and positive sentiment, derived from crew audio communications, plotted against self-reported trust-in-ADS scores at the individual level. Negative sentiment displays a small but statistically significant negative correlation with trust-in-ADS ($r = –0.22$, $p < 0.05$), indicating that higher proportions of negative sentiment are associated with lower trust ratings. Neutral sentiment shows a similarly small but significant positive correlation ($r = 0.22$, $p < 0.05$), suggesting that higher neutral sentiment is linked to higher trust-in-ADS. Positive sentiment shows no meaningful association ($r = –0.02$). Overall, these results suggest that at the individual level, sentiment from audio communications has only a modest relationship with trust-in-ADS, with negative and neutral sentiment showing the most notable (though small) effects.

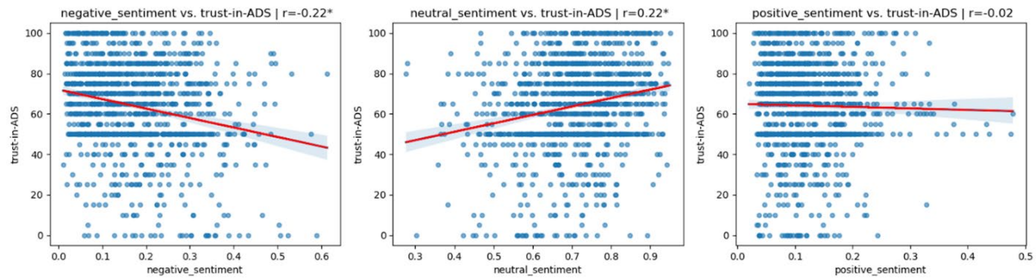

Figure 8. Correlations between sentiment scores features and trust-in-ADS ratings. See Figure 2 for plot layout and description details.

4.2.1.2 ECG-based Features

Figure 9 shows scatter plots of ECG-derived features (HR, RMSSD, and SDSD) plotted against self-reported trust in ADS scores at the individual level. HR shows no meaningful association with trust in ADS (r = –0.05), indicating that average HR is not a strong predictor of trust. RMSSD exhibits a small but statistically significant negative correlation (r = –0.16, $p < 0.05$), suggesting that higher parasympathetic activity and greater HRV are modestly associated with lower trust in ADS. Similarly, SDSD shows a weak but significant negative correlation (r = –0.14, $p < 0.05$), indicating that greater short-term variability between successive *RR* intervals is also linked to slightly lower trust ratings. Although these correlations are statistically significant for RMSSD and SDSD, the effect sizes are small, suggesting limited predictive value of these ECG features for trust in ADS at the individual level.

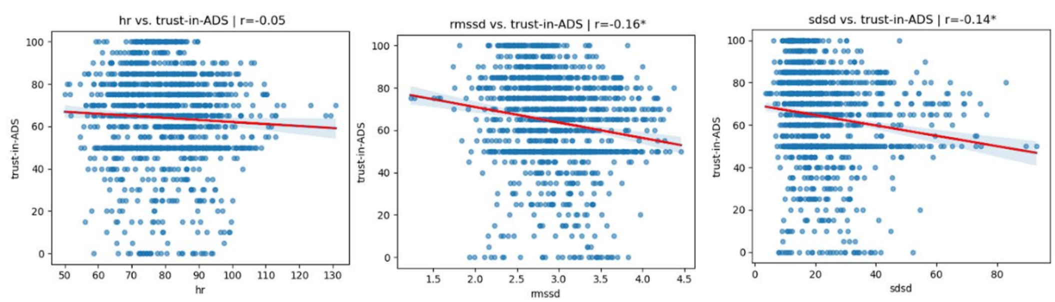

Figure 9. Correlations between ECG-derived features and trust-in-ADS ratings. See Figure 3 for plot layout and description details.

4.2.1.3 Eye-based Features

Figure 10 shows scatter plots of eye-tracking–derived features plotted against self-reported trust in ADS scores at the individual level. Pupil size shows a moderate negative correlation (r = –0.26, $p < 0.05$), indicating that larger pupil sizes are associated with lower trust in ADS. Saccade count (r = –0.22, $p < 0.05$), saccade distance mean (r = –0.23, $p < 0.05$), and saccade velocity mean (r = –0.18, $p < 0.05$)

also show small but statistically significant negative correlations, suggesting that greater saccade activity is linked to lower trust levels. Fixation duration shows a weak positive correlation ($r = 0.06$, $p < 0.05$), while blink count is not significantly associated with trust in ADS ($r = 0.02$). Overall, the results suggest that oculomotor behaviors related to gaze dynamics—particularly saccade metrics and pupil size—are modestly related to trust in ADS, although effect sizes remain too small to moderate.

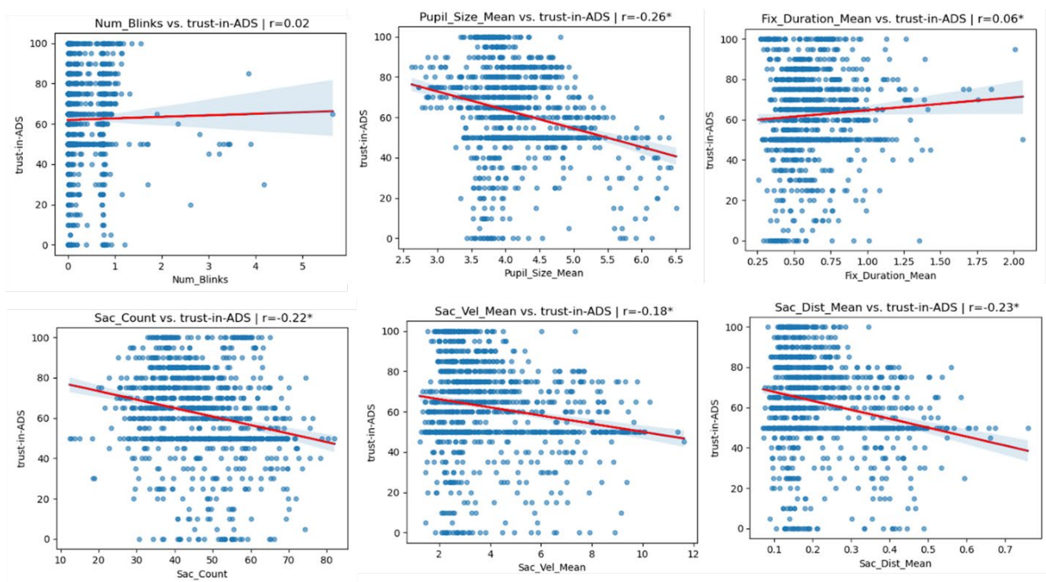

Figure 10. Individual-level correlations between eye-based features and trust-in-ADS ratings. See Figure 4 for plot layout and description details.

4.2.2 Team-Level Correlation Plots

4.2.2.1 Team-Level Audio Communication Features

Figure 11 shows scatter plots of predicted sentiment scores for negative, neutral, and positive sentiment averaged across all crew members within each phase, plotted against team-level trust in ADS ratings. Negative sentiment shows a small but statistically significant positive correlation ($r = 0.27$, $p < 0.05$), indicating that higher average negative sentiment is linked to higher trust in ADS at the team level. In contrast, neutral sentiment displays a small but statistically significant negative correlation ($r = -0.27$, $p < 0.05$), suggesting that higher neutral sentiment is associated with lower trust in ADS. Positive sentiment shows no meaningful relationship ($r = -0.00$). Overall, these results indicate that, when averaged to the team level, certain sentiment proportions are modestly associated with trust in ADS.

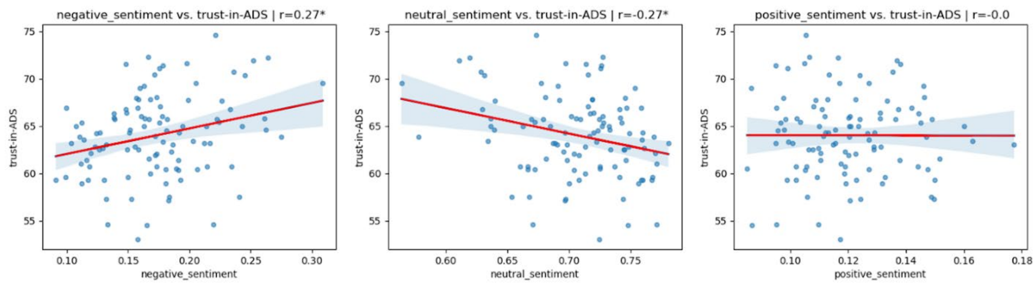

Figure 11. Team-level correlations between communication sentiment features and trust-in-ADS ratings. See Figure 2 for plot layout and description details.

4.2.2.2 Team-Level ECG-based Features

Figure 12 shows scatter plots of team averaged ECG features (HR, RMSSD, and SDSD) plotted against trust in ADS ratings. HR shows a small but statistically significant negative correlation (r = –0.22, p < 0.05), suggesting that teams with higher average HRs tend to report lower trust in ADS. RMSSD shows no meaningful relationship (r = –0.02), indicating that overall HRV at the team level is not strongly associated with trust. Similarly, SDSD shows a negligible correlation (r = –0.04), suggesting little-to-no link between short-term variability in *RR* intervals and trust in ADS. Overall, these findings indicate that at the team level, HR exhibits a modest negative relationship with trust in ADS, while variability-based measures (RMSSD, SDSD) show minimal associations.

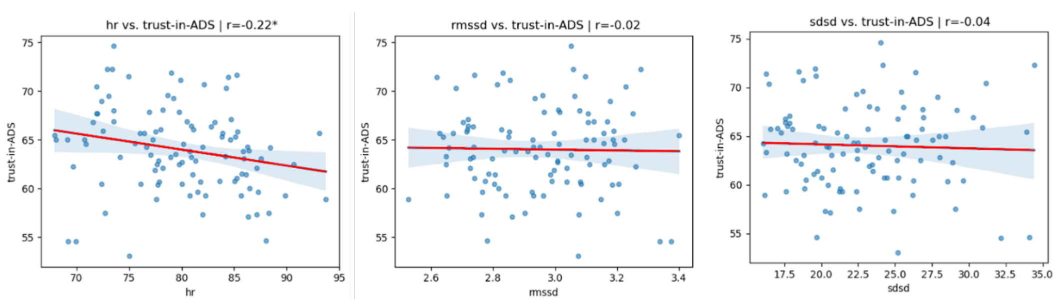

Figure 12. Team-level correlations between ECG-based features and trust in ADS ratings. See Figure 3 for plot layout and description details.

4.2.2.3 Team-Level Eye-based Features

Figure 13 shows scatter plots of team-averaged eye-based features plotted against trust in ADS ratings. Pupil size shows a small but statistically significant negative correlation (r = –0.21, p < 0.05), indicating that teams with larger average pupil sizes tend to report lower trust in ADS. Pupil size standard deviation shows a small positive correlation (r = 0.19), suggesting that greater variability in pupil size may be linked to slightly higher trust. Saccade count displays a small negative correlation (r = –0.19), indicating that teams with more frequent saccades tend to report lower trust in ADS.

Saccade distance means shows a small positive correlation ($r = 0.25$, $p < 0.05$), and this suggests that greater average saccade distances are associated with higher trust in ADS. Similarly, saccade velocity shows moderate positive correlation ($r = 0.3$, $p < 0.05$), indicating that faster average saccade velocity could be linked to greater trust in ADS. Saccade velocity standard deviation shows a weak positive correlation ($r = 0.12$), suggesting only a slight tendency for greater variability in saccade velocity to be associated with higher trust.

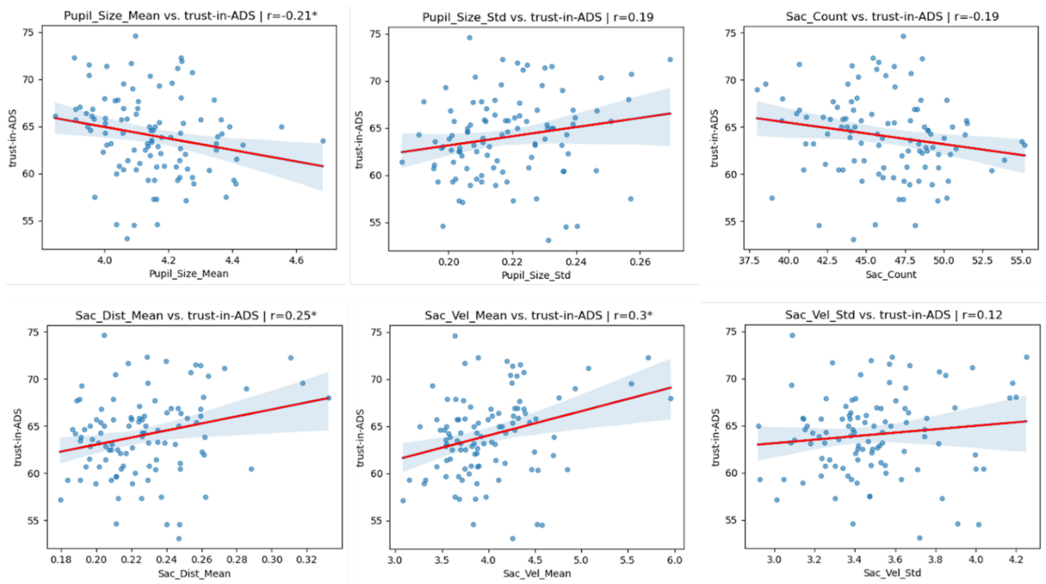

Figure 13. Team-level correlations between eye-based features and trust-in-ADS ratings. See Figure 4 for plot layout and description details.

4.3 Task Cohesion

Here, we report correlations between the mission-level self-reported platoon task cohesion and the multimodal behavioral and physiological features at both the individual and team levels.

4.3.1 Individual-Level Correlation Plots

4.3.1.1 Individual-Level Audio Communication Features

Figure 14 shows scatter plots of predicted sentiment scores for negative, neutral, and positive sentiment, derived from individual audio communications and plotted against self-reported platoon task cohesion ratings. Across all three sentiment types, correlations with cohesion are negligible: negative sentiment ($r = 0.03$), neutral sentiment ($r = -0.02$), and positive sentiment ($r = -0.01$). The scatter plots show broad dispersion of data points across the cohesion scale, with no clear trends or clustering patterns. These results suggest that, at the individual level, sentiment composition in audio communications is not meaningfully related to platoon-task cohesion.

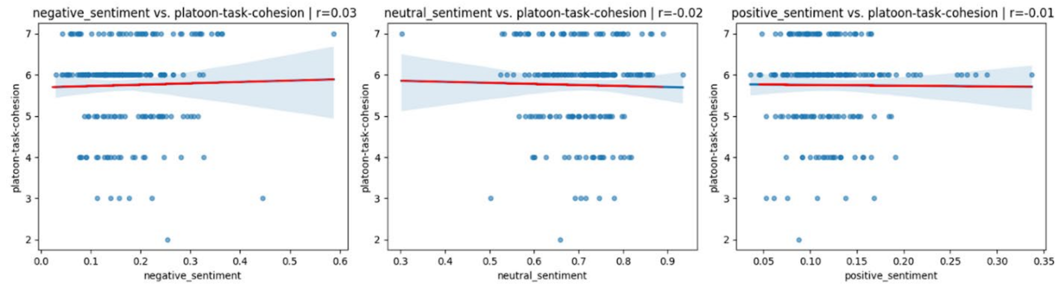

Figure 14. Correlations between sentiment scores features and platoon-task cohesion ratings. See Figure 2 for plot layout and description details.

4.3.1.2 Individual-Level ECG-based Features

Figure 15 shows scatter plots of ECG-derived features (HR, RMSSD, and SDSD), plotted against self-reported platoon-task cohesion ratings at the individual level. HR shows no meaningful association with cohesion (r = 0.03), suggesting that average HR is not a strong indicator of perceived cohesion. RMSSD shows a small negative correlation (r = –0.06), indicating a weak tendency for greater HRV to be linked to lower cohesion ratings. Similarly, SDSD shows a negligible negative correlation (r = –0.02). Overall, these results indicate that individual-level ECG features have minimal predictive value for platoon-task cohesion in this dataset.

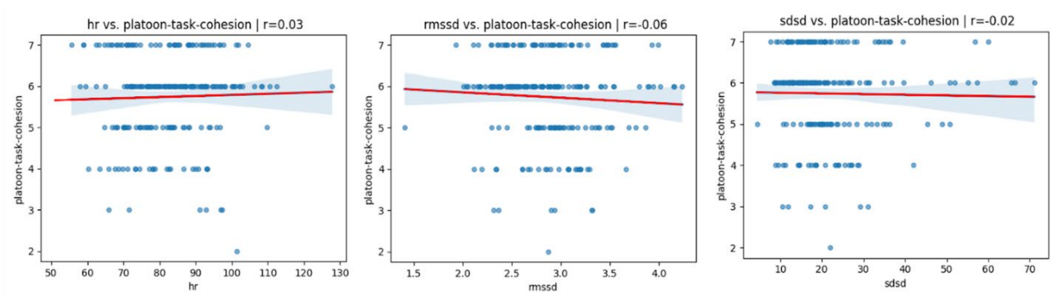

Figure 15. Correlations between ECG-derived features and platoon-task cohesion ratings. See Figure 3 for plot layout and description details.

4.3.1.3 Individual-Level Eye-based Features

Figure 16 shows scatter plots of eye-based features plotted against self-reported platoon-task cohesion ratings at the individual level. Pupil size shows a moderate negative correlation (r = –0.35, $p < 0.05$), indicating that larger average pupil sizes are associated with lower cohesion. Pupil size standard deviation also shows a significant negative correlation (r = –0.26, $p < 0.05$), suggesting that greater variability in pupil size corresponds to lower cohesion ratings. Saccade count shows no meaningful relationship (r = 0.00). However, saccade distance mean (r = –0.32,

p < 0.05), saccade velocity mean (r = –0.21, p < 0.05), and saccade velocity standard deviation (r = –0.21, p < 0.05) all display small-to-moderate negative correlations, indicating that greater saccade distances and higher or more variable saccade velocities tend to be linked with lower cohesion ratings.

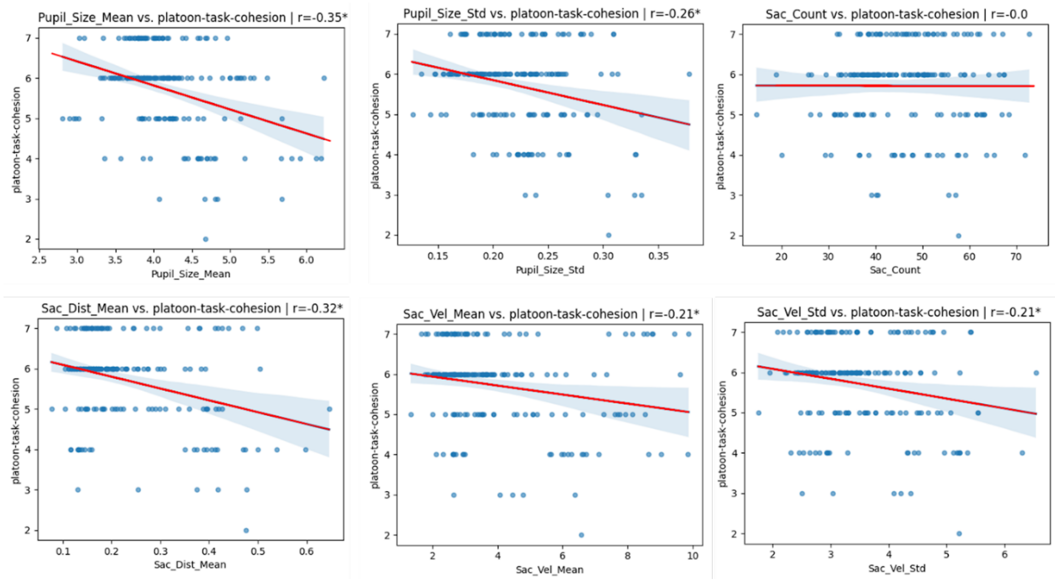

Figure 16. Individual station-level correlations between eye-based features and platoon-task cohesion ratings. See Figure 4 for plot layout and description details.

4.3.2 Team-Level Correlation Plots

4.3.2.1 Team-Level Audio Communication Features

Figure 17 shows the scatter plots of predicted sentiment scores for negative, neutral, and positive sentiment, averaged across all crew stations within each mission and plotted against the platoon-task cohesion ratings. Negative sentiment displays a strong negative correlation with cohesion (r = –0.71, p < 0.05), suggesting that missions with higher proportions of negative sentiment in communication tend to have lower cohesion ratings. Neutral sentiment shows a strong positive correlation (r = 0.67, p < 0.05), and this indicates that higher levels of neutral communication are associated with greater cohesion. Positive sentiment exhibits a weaker positive correlation (r < 0.41), suggesting a modest link between more positive communication and higher cohesion.

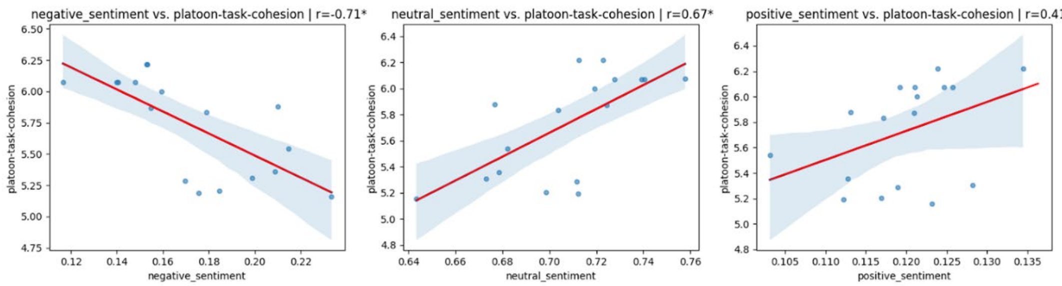

Figure 17. Team-level correlations between communication sentiment features and platoon-task cohesion ratings. See Figure 2 for plot layout and description details.

4.3.2.2 Team-Level ECG-based Features

Figure 18 shows the scatter plots of mission-level averaged for HR, RMSDD, and SDSD plotted against platoon-task cohesion ratings. HR shows a moderate positive correlation with cohesion ($r = 0.50$, $p < 0.05$), indicating that missions with higher average HRs tended to receive higher cohesion ratings. Both RMSDD and SDSD show small positive association, which is not statistically significant; thus, inducing limited predictive value.

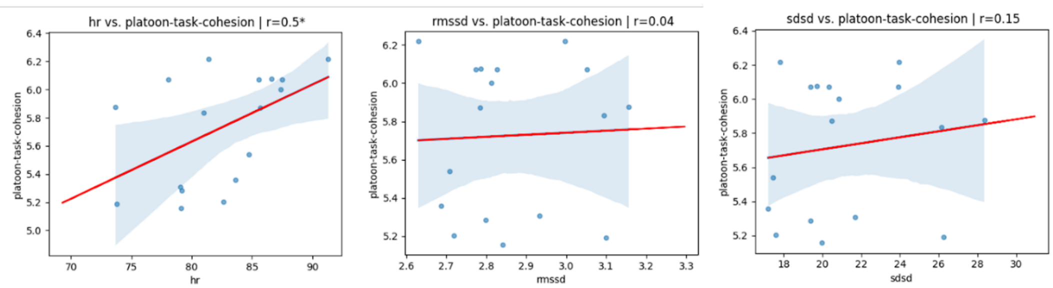

Figure 18. Team-level correlations between ECG-based features and platoon-task cohesion ratings. See Figure 3 for plot layout and description details.

4.3.2.3 Team-Level Eye-based Features

Figure 19 presents mission-level correlations between eye-based features and platoon-task cohesion ratings. Among the examined metrics, mean pupil size ($r = 0.56$, $p < 0.05$) and saccade count ($r = 0.69$, $p < 0.05$) showed strong positive relationships with cohesion, indicating that larger average pupil diameters and higher saccade frequencies were associated with higher perceived task cohesion at the platoon level. In contrast, pupil size variability ($r = –0.78$, $p < 0.05$), mean saccade distance ($r = –0.75$, $p < 0.05$), mean saccade velocity ($r = –0.72$, $p < 0.05$), and saccade velocity variability ($r = –0.56$, $p < 0.05$) exhibited strong negative correlations, suggesting that greater variability and higher movement amplitudes or speeds were linked to lower cohesion ratings. These findings highlight that both the magnitude and stability of eye-movement patterns are relevant indicators of group-level coordination and cohesion.

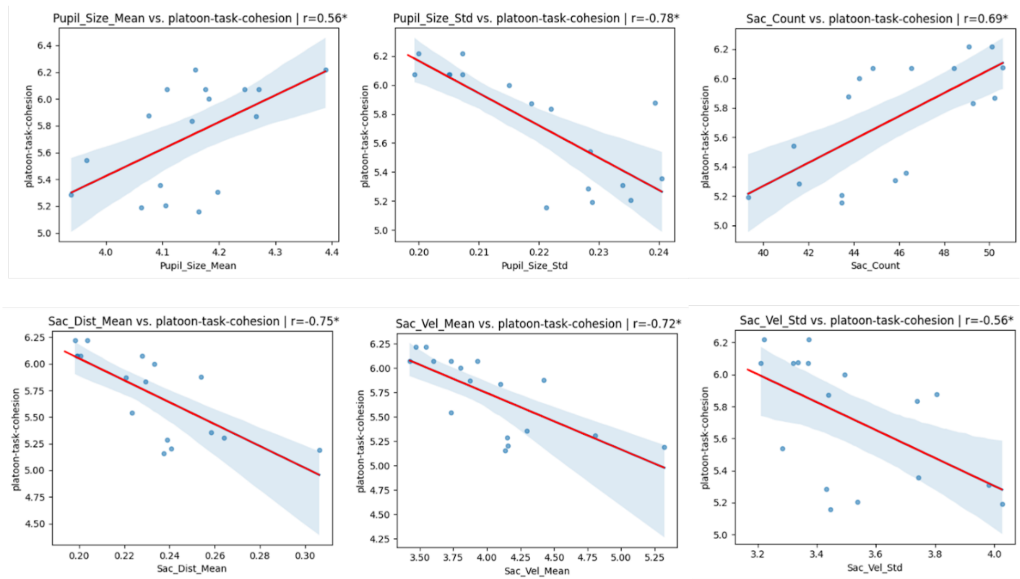

Figure 19. Team-level correlations between eye-based features and platoon-task cohesion ratings. See Figure 4 for plot layout and description details.

4.4 Trust and Cohesion Models Development and Validation

Extending beyond the correlation analysis conducted in the previous sections, we developed predictive models to estimate individual team trust and task cohesion with the features generated from the multimodal physiological and behavioral data. The goal is to enable continuous assessment of team states that could facilitate near real-time intervention to improve team performance. We leveraged the open-sourced automated machine learning (AutoML) library called AutoGluon to build models of team trust and cohesion. AutoGluon is a state-of-the-art AutoML framework specializing in hyperparameter optimization and multilayer stacking that combines both the traditional machine learning models and deep learning techniques to ensure models' robustness and accuracy. The AutoGluon Python package fits a variety of models that include gradient boosting, random forests, K-nearest neighbors, and neural network approaches. It then fits a weighted ensemble method that combines predictions from each model and stacks them in layers to achieve superior accuracy. This ensemble approach is used as the final model for predictions. Details about each modeling approach can be found in Erickson et al.[18]

The pipeline for training the models with AutoGluon involved having the data in a tabular format, where rows represent separate instances and columns contain distinct features. For our dataset, we have a total of 1345 rows (combining across the 2 groups, 13 missions, 14 crew, and 4 phases) and 12 features generated from the audio communication data, ECG, and eye-tracking data. Those features are positive sentiment, neutral sentiment, negative sentiment, HR, RMSSD, SDSD, mean and standard deviation of the pupil size, saccade count, saccade distribution mean, and mean and standard deviation of the saccade velocity (see details in previous sections).

A tenfold cross-subject validation was performed where the dataset was split by individual crew into 10 parts (8 parts with 3 crew + 2 parts with 2 crew = 28 crew). The models were trained 10 times, with each iteration using 1 part (data from 2 to 3 crew) as the test set and the remaining 9 parts for training. This ensures model reliable performance and better generalization. Models were trained separately for the three self-reported trust and cohesion metrics. Figures 20–22 show model predictions for the trust-in-team, trust-in-ADS, and platoon-task cohesion, respectively.

We evaluated model performance by two metrics: the correlation coefficient (*r*) computed from a linear fit and the root mean square error (*rmse*) between the predicted and true values.

Correlation coefficient measures the relationship between the predicted and true values. It ranges from –1 for a perfectly inverse relationship to 1 for a perfectly positive correlation. *R* values at or close to zero indicate no or weak linear relationship. The *rmse* measures the averaged difference between the model's predicted values and the actual values. The lower the *rmse* value the better the model in its predictions that are close to the true values.

Figure 20 shows a strong and significant positive correlation between the predicted and the true ratings of team trust. The values of the model predictions were between 4.0 and 7.0, reflecting the highly skewed distribution of the data in which most of the individual crew responses to the questionnaire are in the high trust region (>4.0). For the trust-in-team model, the *rmse* value is 1.08, indicating the predicted values differ from the actual values by about 1.0 point on average. On a scale from 1 to 7, that translates to a less than 15% difference between the predicted and the true trust in team ratings.

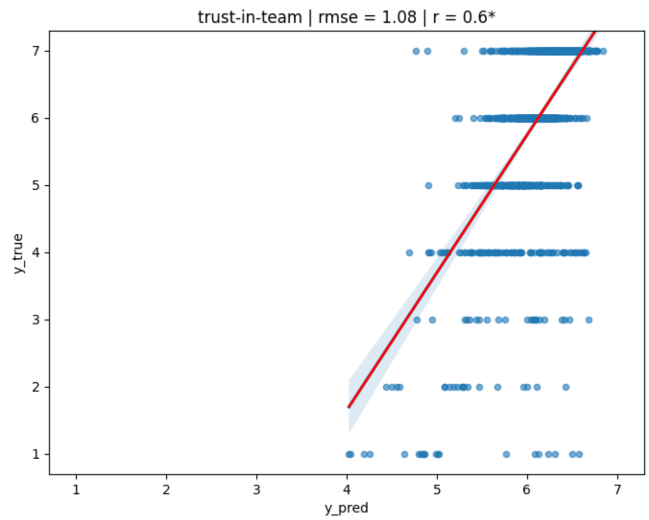

Figure 20. Correlation between the model predictions (x-axis) and the true value (y-axis) of trust-in-team. Trust-in-team model performance was evaluated by the *rmse* and *r* values with significant levels indicated by the asterisk.

Model predicting trust in the ADS system performs well with a higher correlation at significant levels between the predicted and true ratings (Figure 21). The spread of the predictions are closely aligned with the true values, suggesting a better fit and generalization for the model. The *rmse* value of 14.3 indicates the averaged percent difference between the predicted and true trust-in-ADS ratings, comparable to the trust-in-team model performance in that regard.

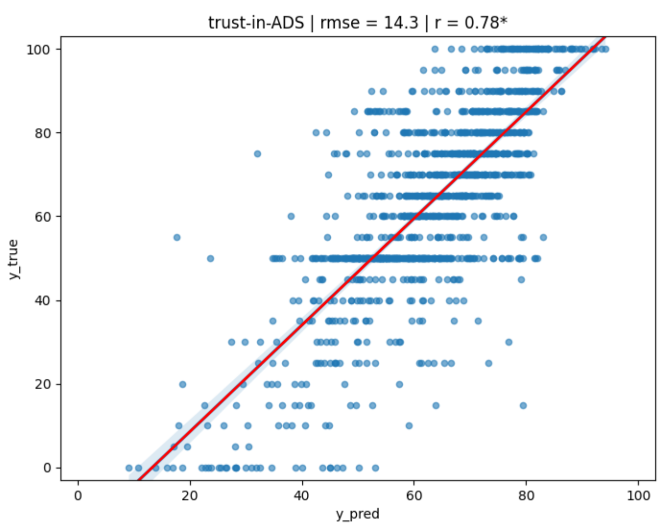

Figure 21. Correlation between the model predictions (x-axis) and the true value (y-axis) of trust-in-ADS. Trust-in-ADS model performance was evaluated by the *rmse* and *r* values with significant levels indicated by the asterisk.

The platoon-task cohesion model did not perform as well as the other two models (Figure 22). The predicted scores and the true ratings are weakly correlated ($r = 0.1$*), yet significant. Similar to the team trust model, the platoon-task cohesion model predictions mostly fall between 4.0 and 7.0. As discussed previously, this reflects the highly skewed data used in training the model in which most ratings are greater than 4.0. The *rmse* is comparable to what was seen in the team-trust model where the averaged difference between the predicted and true values are within the 15% range.

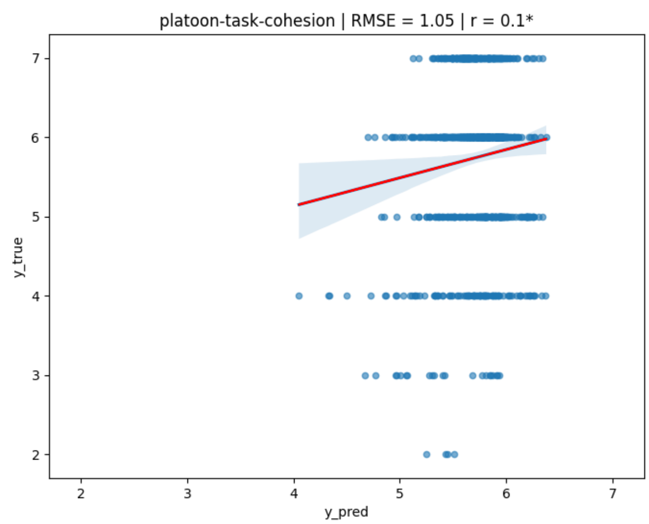

Figure 22. Correlation between the model predictions (x-axis) and the true value (y-axis) of platoon-task cohesion. Platoon-task cohesion model performance was evaluated by the *rmse* and *r* values with significant levels indicated by the asterisk.

5. Discussion

These analyses examined the relationships among multimodal behavioral and physiological features, derived from audio communication, ECG, and eye-tracking data, and three key outcome measures in team-based simulated missions: trust-in-team, trust-in-ADS, and platoon-task cohesion. By analyzing both individual-level and aggregated-level (team or mission) correlations, we identified consistent trends in how sentiment, cardiac activity, and oculomotor behavior are associated with these psychosocial constructs.

5.1 Audio Communication and Sentiment

Across outcomes, sentiment features from crew audio communications exhibited stronger correlations at aggregated levels (team or mission) than at the individual level. For example, negative sentiment was a strong negative predictor of trust-in-team (r = –0.55) and cohesion (r = –0.71) at aggregated levels, while neutral sentiment was positively related to these outcomes (r = 0.48 for trust-in-team; r = 0.67 for cohesion). Positive sentiment was more weakly related, showing modest or negligible associations.

By contrast, at the individual level, sentiment–outcome relationships were weaker, often negligible for trust-in-team and cohesion; however, small negative correlations were observed between negative sentiment and trust in ADS. These findings suggest that sentiment composition in communications is more predictive of collective perceptions when measured at the group scale; this is likely because aggregation reduces individual noise, suppresses outliers, and captures emergent team communication patterns. This aligns with prior work showing that communication tone is important for group cohesion and trust formation in high-stakes environments.[19]

Interestingly, the results for trust-in-ADS diverged between individual and team levels. At the individual level, negative sentiment displayed a small but statistically significant negative correlation with trust-in-ADS (r = –0.22, $p < 0.05$), suggesting that when individuals expressed more negative affect in their communications, they tended to report lower trust in the ADS. However, when sentiment was aggregated to the team level, the relationship reversed: negative sentiment showed a small positive correlation with trust (r = 0.27, $p < 0.05$).

This contrast highlights how the meaning of sentiment signals can shift with scale. At the individual level, a negative tone may reflect personal frustration or lack of confidence in the ADS. Yet, at the team level, negative sentiment may represent critical engagement and vigilant monitoring, where crews collectively articulate concerns, challenge the system, and in doing so develop greater shared confidence in its reliability. This finding underscores the importance of contextualizing both the level of analysis and the target of trust, whether interpersonal (team) or technological (ADS), as the same affective cues may carry opposite implications across scales.

5.2 ECG-derived Cardiac Activity

ECG-derived features revealed small-to-moderate associations with trust and cohesion, though the magnitude and direction of effects varied by both the outcome examined and the level of aggregation.

HR consistently showed stronger associations at aggregated levels than at the individual level. For trust-in-team, HR was positively correlated at both scales, but effects were much stronger at the team level (r = 0.54) compared to the individual level (r = 0.13). Similarly, HR was positively correlated with cohesion at the mission level (r = 0.50), whereas at the individual level the association was negligible (r = 0.03). These findings suggest that elevated collective arousal may serve as a marker of coordinated engagement and alignment, reinforcing perceptions of both trust in teammates and task cohesion. In contrast, HR showed a different pattern for trust-in-ADS; at the individual level it was not significantly related (r = –0.05), while at the team level it was negatively correlated (r = –0.22, $p < 0.05$). This divergence highlights that while HR appears to support interpersonal trust and group cohesion, it may reflect discomfort, workload strain, or heightened vigilance in the context of evaluating technology.

HR variability metrics (RMSSD and SDSD) showed more consistent negative relationships with trust, particularly for trust-in-team and trust-in-ADS. At the individual level, RMSSD (r = –0.12 to –0.16) and SDSD (r = –0.08 to –0.14) were small but reliably negative predictors of trust. At the team level, these relationships became more pronounced; RMSSD (r = –0.25) and SDSD (r = –0.23) both showed moderate negative correlations with trust-in-team, suggesting that lower HRV at the collective scale aligns with stronger shared trust. One interpretation is that reduced variability reflects synchronized arousal and attentional alignment, whereas greater variability may signal relaxation or disengagement, corresponding to weaker trust in peers or the ADS.

For cohesion, however, HRV effects differed. At the individual level, RMSSD (r = –0.06) and SDSD (r = –0.02) showed negligible negative relationships. At the mission level, both variability measures were weakly positive, though not statistically significant. This suggests that variability may not be a reliable marker of perceived cohesion, even if it relates more directly to trust.

Together, these results underscore that the arousal–trust relationship is not uniform. Elevated HR at the team or mission level appears to support collective perceptions of trust and cohesion, consistent with the idea that shared physiological activation signals engagement and alignment during coordinated performance. By contrast, HRV is generally negatively related to trust—particularly for team trust—but less meaningful for cohesion. Finally, the divergence between HR's positive link to interpersonal trust and cohesion versus its negative link to trust-in-ADS underscores the importance of considering the target of trust (human vs. technological) when interpreting physiological markers.

5.3 Eye-Tracking and Oculomotor Behavior

Eye-based metrics showed some of the strongest and most consistent associations with trust and cohesion, particularly at the aggregated team level. While individual-level correlations were often small, averaging across crew stations revealed clearer patterns linking oculomotor dynamics to collective perceptions.

At the individual station level, mean pupil size (r = –0.22), mean saccade velocity (r = –0.26), and mean saccade distance (r = –0.35) were all modest negative predictors of trust, while blink count, fixation duration, and saccade count showed negligible effects. In contrast, team-level analyses revealed stronger and more interpretable relationships; mean pupil size was positively associated with trust (r = 0.42), whereas pupil size variability was strongly negatively correlated (r = –0.64). Similarly, saccade count was positively associated with trust (r = 0.52), but saccade distance (r = –0.60), saccade velocity (r = –0.58), and velocity variability (r = –0.49) were strongly negative predictors. Taken together, these results suggest that stable and frequent gaze activity supports stronger team trust, while erratic or extreme eye movements signal reduced confidence in peers.

At the individual level, larger pupil size (r = –0.26), more frequent saccades (r = –0.22), longer saccade distances (r = –0.23), and faster saccade velocities (r = –0.18) were all modest negative predictors of trust-in-ADS, while fixation duration showed a weak positive relationship (r = 0.06). At the team level, however, the pattern diverged: mean pupil size was slightly negatively associated with trust (r = –0.21), but pupil size variability was positively related (r = 0.19). Moreover, saccade distance (r = 0.25) and saccade velocity (r = 0.30) became positive predictors of trust-in-ADS. This contrast highlights that individual oculomotor signals of uncertainty (e.g., large pupils, frequent saccades) may aggregate into group-level indicators of collective vigilance and engaged monitoring, which in turn can foster higher trust in the ADS.

For cohesion, individual-level effects were broadly negative; mean pupil size (r = –0.35), pupil size variability (r = –0.26), saccade distance (r = –0.32), mean saccade velocity (r = –0.21), and velocity variability (r = –0.21) all predicted lower ratings. At the mission level, however, the pattern reversed for some features; mean pupil size (r = 0.56) and saccade count (r = 0.69) were strong positive predictors of cohesion, while variability and amplitude/velocity measures were strongly negative (pupil size variability $r = -0.78$, saccade distance $r = -0.75$, saccade velocity $r = -0.72$, velocity variability $r = -0.56$). These findings suggest that stable but active visual engagement (larger pupils, frequent saccades) is a marker of stronger group cohesion, whereas variability and extreme movements may reflect instability, distraction, or lack of alignment.

5.4 Trust and Cohesion Models

Results in Section 5.3 show our machine learning models are capable of predicting trust and cohesion at the individual level with good accuracy using their physiological and behavioral data. For team trust and trust-in-ADS, the model predictions are within 15% of the true ratings; and the two values are highly correlated at significant levels. For the platoon-task cohesion model, the predictions are marginally correlated with the true ratings, yet significant; and the predicted scores differ from the true ratings in the 15% range. The model development and validation improved significantly in that we have four times the self-reported trust and cohesion responses (the questionnaire was administered more frequently—i.e., at the end of each phase as opposed to at the end of each mission). There are four phases per mission and thus more labeled data compared to previous years.[20,21] Our trust and cohesion models developed using features extracted from multimodal physiological and behavioral signals collected in near–real-time would enable continuous predictions of trust-in-team, trust-in-technology, and platoon-task cohesion. This continuous monitoring and assessment of team states from individual crew members has the potential to provide just-in-time intervention needed to improve team performance and mission outcome.

6. Conclusions

This study examined multimodal physiological and behavioral features, audio sentiment, ECG-derived cardiac activity, and eye-tracking metrics, in relation to trust-in-teammates, trust-in-ADS, and platoon-task cohesion, across individual, team, and mission levels. Several consistent patterns emerged.

First, aggregation strengthened predictive relationships. At the individual level, correlations between features and outcomes were generally small or negligible, reflecting the variability of individual responses. However, at the team and mission levels, aggregation suppressed outliers and amplified group-level dynamics. For example, neutral sentiment in team communications was a strong positive predictor of both trust-in-team and cohesion, while negative sentiment was strongly detrimental to cohesion at the mission level. Similarly, team-averaged HR showed a strong positive association with trust in the team, whereas at the individual level the effect was minimal. These findings highlight the importance of scale in interpreting multimodal signals of trust and cohesion.

Second, physiological arousal and variability displayed divergent roles depending on the target of trust. Elevated HR at the group level aligned with higher trust-in-teammates and greater cohesion, suggesting that synchronized physiological arousal may serve as a marker of engagement and alignment. In contrast, higher HR was negatively associated with trust-in-ADS, pointing to the possibility that physiological arousal in this context reflects heightened vigilance or discomfort with technology. Conversely, HRV measures (RMSSD, SDSD) tended to show negative associations with trust, particularly at aggregated levels, suggesting that reduced variability and more synchronized cardiac activity may underpin shared perceptions of reliability within teams.

Third, eye-based metrics provided some of the strongest and most consistent predictors, especially when aggregated. Stable gaze activity (larger average pupil size, frequent saccades) supported higher cohesion and trust-in-team, while variability and extreme gaze behaviors (large saccade distances, high saccade velocity) were linked to reduced trust and cohesion. Interestingly, for trust-in-ADS, these relationships often reversed at the group level; eye dynamics that signaled individual uncertainty appeared to reflect collective vigilance when averaged, aligning with higher trust-in-ADS. This contrast underscores how the meaning of behavioral cues shifts with context and level of analysis.

Finally, predictive modeling results confirmed the feasibility of near real-time trust and cohesion estimation. Using AutoGluon's ensemble approach with multimodal features, the trust-in-team and trust-in-ADS models achieved strong predictive accuracy, with predictions within approximately 15% of true ratings. While the cohesion model performed less strongly, its predictions were still statistically significant and comparable in error margins. These results were enabled by a richer dataset with more frequent self-reports than prior studies, providing a stronger foundation for model training and validation. The cohesion model is being integrated into the Human–Autonomy Teaming Trust Toolkit.[11,12]

Overall, the findings suggest that multimodal monitoring of team states is both possible and informative. Physiological and behavioral markers, particularly when aggregated, provide meaningful signals of trust and cohesion, with distinct patterns emerging for interpersonal versus technological trust. Importantly, continuous prediction models developed here can support just-in-time interventions, helping teams maintain alignment, adapt to challenges, and ultimately improve mission performance. Looking ahead, future work should expand data collection across more teams and operational contexts, refine and externally validate predictive models with advanced algorithms and explainability methods, and transition from offline analyses to real-time monitoring within training and operational environments. Together, these next steps would enable the transition from experimental validation to robust, field-ready systems capable of supporting adaptive human–autonomy teaming.

7. References

1. Grossman R, Feitosa J. Team trust: a review of the literature and a new conceptual model. The Oxford Handbook of Teamwork and Collaborative Work. University of Oxford (UK); 2018. p. 395–420.

2. Grossman R et al. The relationship between team trust and team effectiveness: a meta-analytic review. Hum Factors. 2018;60(5):591–604.

3. Drnec K, Metcalfe J. Trust dynamics in human-robot teams. Proc Hum Factors Ergon Soc Annu Meet. 2016;60(1):1359–1363.

4. Hancock PA et al. A meta-analysis of factors affecting trust in human-robot interaction. Hum Factors. 2011; 53(5):517–527.

5. Merritt SM et al. A review of the literature on trust in automation. Theor Issues Ergon Sci. 2013;14(3):282–301.

6. Dion KL. Group cohesion: from "field of forces" to multidimensional construct. Group Dyn Theory Res Pract. 2000;4(1):7.

7. Carron AV. Cohesiveness in sport groups: interpretations and considerations. J Sport Psychol. 1982;4(2):123–138.

8. Salas E et al. The science of teamwork: progress, reflections, and the road ahead. Am Psychol. 2015;70(6):554.

9. Carless SA, De Paola C. The measurement of cohesion in work teams. Small Group Res. 2000;31(1):71–88.

10. Krausman A et al. Trust measurement in human–autonomy teams: development of a conceptual toolkit. DEVCOM Army Research Laboratory (US); 2022. Report No.: ARL-TR-9461.

11. Neubauer C et al. Human–autonomy teaming trust toolkit (HAT^3) executive summary. DEVCOM Army Research Laboratory (US); 2022. Report No.: ARL-TR-9622.

12. Neubauer C et al. Human–autonomy teaming trust toolkit (HAT^3) software development documentation and user guide. DEVCOM Army Research Laboratory (US); 2022. Report No.: ARL-TR-9623.

13. Gremillion GM et al. Technologies for automated dynamic resource allocation in the Next Generation Combat Vehicle. DEVCOM Army Research Laboratory (US); 2024. Report No.: ARL-TN-1234.

14. Craig TY, Kelly JR. Group cohesiveness and creative performance. Group Dyn Theor Res Pract. 1999;3(4);243.

15. Kucukosmanoglu M et al. Exploring trust in AI-supported military teams using sentiment analysis. Interaction Studies. Forthcoming 2026.

16. Dalmaijer ES, Mathôt S, Van der Stigchel S. PyGaze: an open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. Behav Res Methods. 2014:46(4);913–921.

17. Mahanama B et al. Eye movement and pupil measures: a review. Front Comput Sci. 2022;3:733531.

18. Erickson N et al. AutoGluon-tabular: robust and accurate automl for structured data. arXiv. 2020. arXiv:2003.06505.

19. Baker AL et al. Approaches for assessing communication in human-autonomy teams. Hum–Intell Syst Integr. 2021;3(2):99–128.

20. Crowell HP et al. Potential indicators of team trust measured in a Next Generation Combat Vehicle simulator. DEVCOM Army Research Laboratory (US); 2023 Feb. Report No.: ARL-TR-9647. p. 56.

21. Crowell HP et al. Metrics of team trust and team cohesion measured during an experiment examining adaptive situation awareness technologies in a Next Generation Combat Vehicle simulator. DEVCOM Army Research Laboratory (US); Forthcoming 2026.

List of Symbols, Abbreviations, and Acronyms

ADS	Automatic Detection System
AI	artificial intelligence
ARL	Army Research Laboratory
AutoML	automated machine learning
BLUFOR	friendly forces
DEVCOM	U.S. Army Combat Capabilities Development Command
ECG	electrocardiogram
HICON	Higher Control
HR	heart rate
HRV	heart rate variability
INFORMS	Information for Mixed Squads (laboratory)
MCV	manned combat vehicle
MTRS	modular-trigger response system
NGCV	Next Generation Combat Vehicle
OPFOR	opposing force
r	correlation efficient
RCV	robotic combat vehicle
rmse	root mean square error
RMSSD	root mean square of successive differences
SC3	Scalable Cross-Echelon Command and Control
SDSD	standard deviation of the differences between adjacent *RR* intervals
Std	standard

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLB CI TECH LIB