Nonlinear Dynamics, Psychology, and Life Sciences, Vol. 29, No. 4, pp. 529-552.

Bio-behavioral Team Dynamics Measurement System: Multimodal Sensing, Dynamical Systems Modeling, and Machine Learning Pipelines to Predict and Characterize Team Performance

Garima Arya Yadav[1], *Arizona State University, Tempe AZ,* **Bethany K. Bracken**, *Charles River Analytics, Cambridge, MA*, **Nancy J. Cooke**, *Arizona State University, Polytechnic Campus, Mesa AZ*, **Phillip Desrochers**, *Charles River Analytics, Cambridge, MA*, **Jamie C. Gorman, David A. P. Grimm, Lixiao Huang**, *Arizona State University, Polytechnic Campus, Mesa AZ*, **Molly Kilcullen**, *The Johns Hopkins University, Baltimore, MD*, **Mengyao Li, Emanuel Rojas**, *Georgia Tech, Atlanta, GA*, **Michael Rosen,** *The Johns Hopkins University, Baltimore, MD*, **Matthew J. Scalia**, *Arizona State University, Polytechnic Campus, Mesa AZ*, **Aaron Winder**, *Charles River Analytics, Cambridge, MA*, **Xiaoyun Yin, Elmira Zahmat Doost,** and **Shiwen Zhou,** *Arizona State University, Polytechnic Campus, Mesa AZ*

Abstract: *The DARPA OP TEMPO program seeks to accelerate warfighter readiness by supplying instructors with objective, automatic assessments of team performance during simulation training. To that end, we created the Bio-behavioral Team Dynamics Measurement System (BioTDMS), a multimodal sensing and analytics pipeline that discovers bio-behavioral "signatures" emanating from within the human body and through team-member interactions that predict team performance. BioTDMS employs a layered symbolic dynamics model that converts time-aligned neural, cardio-respiratory, eye tracking, and verbal data, collected using a multimodal sensor suite. Moving-window entropy and mutual information computed across the symbolic sensor space yield real-time metrics that quantify team adaptability following perturbation (e.g., "training injects") and distribution of team members' influence across biological and behavioral subsystems. These features feed a multitask, multi-kernel learning engine that refines performance prediction while preserving explainability through team construct mapping and a command-line user interface. We present preliminary results from field testing a full physical and computational implementation of BioTDMS during Fire Support Team (FiST) training exercises at the U.S. Marine Corps Air-Ground Combat Center, Twentynine Palms, CA. An onsite team instrumented five-person FiST crews with multimodal sensor suites.*

[1] Correspondence address: Garima (Arya) Yadav, 1111 E Apache Blvd Apt 210, Tempe, AZ 85281. E-mail: arya.yadav@asu.edu

Sensor data were processed by BioTDMS for real-time and post hoc analytics. BioTDMS currently accounts for 90 % of variance in a subjective team performance assessment made by instructors, with improvements expected upon further refinements of BioTDMS modeling components. These findings demonstrate BioTDMS's potential as an operational tool for automatic, objective team assessments. Future assessments within air combat teams, including configurations with human-autonomy teaming, will evaluate the generalizability of BioTDMS.

***Key Words*:** bio-behavioral sensing, team coordination, reorganization entropy, multimodal data integration, symbolic dynamics, team performance assessment

INTRODUCTION

In military operations, teamwork is critical for mission success, requiring seamless communication, coordination, and adaptive responsiveness to unpredictable and rapidly evolving situations. Traditionally, instructors assess team performance subjectively, relying heavily on observational judgment and qualitative evaluations, which vary significantly based on instructor expertise, bias, situational awareness, and fatigue. These subjective assessments, while valuable, often lack consistency, precision, and scalability, limiting their effectiveness in reliably evaluating and training warfighters, particularly in large-scale simulations or operations with complex dynamics.

Simulation training provides opportunities for team skill acquisition and maintenance, but methods for objectively quantifying team performance currently rely on subjective, often *post hoc*, assessments that do not scale with increasing training demands for warfighter readiness. Many team training assessments are limited by their summative nature, rather than dynamic assessments across multiple timescales and levels of bio-behavioral analysis. Valid and generalizable methods for objective team assessment are needed.

To address these challenges, the Defense Advanced Research Projects Agency (DARPA) initiated Objective Prediction of Team Effectiveness via Models of Performance Outcomes (OP TEMPO). The overarching goal of OP TEMPO is to develop technology capable of automatically, objectively, and rapidly assessing team performance based on quantifiable physiological, cognitive, and behavioral signatures, comparable to expert observer (trainer) ratings. The rationale underpinning this initiative is clear: objective assessments not only reduce the variability inherent in subjective evaluations but also enable rapid, actionable insights that can inform real-time interventions, training adaptations, and ultimately improve warfighter readiness and operational effectiveness.

Teams are inherently nonlinear: Small changes in one member's cognitive or physiological state can produce disproportionate effects on others, and on collective performance, depending on initial conditions. Prior approaches have often relied on static, unimodal summaries (e.g., average heart rate or total speaking time), which miss the time-dependent, cross-member reorganization that

characterizes effective teamwork. Under repeated perturbation, these dynamics become nonstationary, and approaches are needed to track cross-member reorganization in real time (Gorman & Wiltshire, 2024). This creates a gap between what is typically measured (summative snapshots) and what matters operationally (moment-to-moment adaptation and coordination).

Earlier work on team coordination dynamics (e.g., Gorman, Amazeen, & Cooke, 2010) showed that functional performance emerges from coupled patterns among teammates, while layered dynamics (Gorman, Demir, Cooke, & Grimm, 2019) demonstrated that bio-behavioral reorganization can be organized across interacting layers (e.g., neural, autonomic, communicative). More recently, Gorman and Wiltshire (2024) articulated how increasing dynamic complexity demands multiscale, multilevel analyses. They suggested that in dynamic environments, teams continuously reorganize and adapt, whose nonstationary dynamics can best be assessed using moving window (e.g., entropy; Stevens, Galloway, Gorman, Willemsen-Dunlap, & Halpin, 2012) approaches. We extend these ideas by operationalizing layered, moving-window metrics over synchronized multimodal signals in live training, linking them directly to instructor ratings and predictive models.

In this article we apply a symbolic dynamics framework that (a) discretizes continuous multimodal signals into interactive states, (b) computes reorganization entropy to quantify adaptive exploration of the state space, and (c) uses mutual information to estimate coupling among teammates. These dynamics map onto well-studied team constructs, e.g., real-time team cognition and adaptive coordination (Gorman et al., 2020, 2025) and communicative influence (Reitman, Harrison, Gorman, Lieber, & D'Mello, 2025), providing theory-grounded, interpretable metrics for training contexts.

In response to these objectives, we developed the Bio-behavioral Team Dynamics Measurement System (BioTDMS). BioTDMS integrates multiple sensor modalities, sophisticated data acquisition methods, real-time dynamical systems modeling, and machine learning techniques to provide precise, objective measures of team performance. Unlike previous approaches that typically rely on single-sensor or unimodal assessments, BioTDMS leverages multimodal integration (e.g., EEG, fNIRS, EKG, respiration, eye-tracking, and communication data), significantly enhancing both the reliability and explanatory power of performance predictions that are not constrained by the subjective nature of performance ratings or shared mental models.

A key innovation of BioTDMS is the team coordination dynamics (Gorman et al., 2010, Gorman & Wiltshire, 2024) computer, which operationalizes advanced computational frameworks, including symbolic dynamics modeling (Gorman et al., 2019) and entropy-based reorganization metrics (Gorman et al., 2025, Stevens et al., 2016). By employing a layered symbolic dynamics approach across neurological, autonomic, and communicative layers, the system explicitly differentiates between distinct yet interconnected domains of team functioning, providing insight into how individual and collective

bio-behavioral dynamics interact to impact overall performance outcomes. BioTDMS uses these innovations to translate raw physiological and behavioral data into meaningful indicators of team adaptation, influence, and physiological synchrony.

In this paper, we provide an overview of BioTDMS, describing its technical components, data processing pipelines, predictive analytics, and ontology-based explanatory tools. We describe ongoing validation efforts through field tests conducted during Fire Support Team (FiST) exercises at the U.S. Marine Corps Air-Ground Combat Center, Twentynine Palms, CA, demonstrating the predictive power of BioTDMS for a range of ground truth performance assessments made by their instructors. Here, "ground truth" refers to instructor-provided FiST performance ratings collected on standardized rubrics at the end of each scenario; these ratings serve as the criterion variables for BioTDMS performance prediction. Figure 1 provides an illustration of the BioTDMS data processing and modeling pipelines, including: Multimodal Sensor Suite (MMSS); Bio-behavioral Team Coordination Dynamics Computer (B-TCDC); Performance Prediction and Classification Model (PPCM); and Explanatory Mapping Module (EMM). The following sections provide details of each component, followed by performance prediction validation results from testing at Twentynine Palms.

METHOD

Participants

Four Marine Fire Support Teams (FiSTs) participated; each team comprised five roles: Lead, Joint Terminal Attack Controller (JTAC), Fire Support Officer (FSO), Fire Observer Assistant (FOA), and Fire Operations Center Marine (FOM). Across data collection, we recorded eight scenario runs (≈ 60 min each) across four days, with 2 sessions each day. Four FiST teams participated across eight scenario runs across four days, with two teams participating on the first two days, and two teams participating on the second two days. There were three types of scenarios, and each progressed in difficulty. We consented participants and executed data collection in accordance with IRB and HRPO/OHARO/HRPP approved protocols.

Procedure

At the beginning of each day, all participants were outfitted with ABM B-Alert EEG caps, which they wore for the remainder of the day. Each session followed a common timeline: equipment donning and signal checks (30 min or less); role-specific brief; scenario execution with planned "training injects" (perturbations) to induce adaptation (approx. 60 min.); and equipment doffing (approx. 5 min.). Instructors observed and later completed the ground truth performance ratings per scenario (see below). The details of BioTDMS and different components used are in the following sections in chronological order.

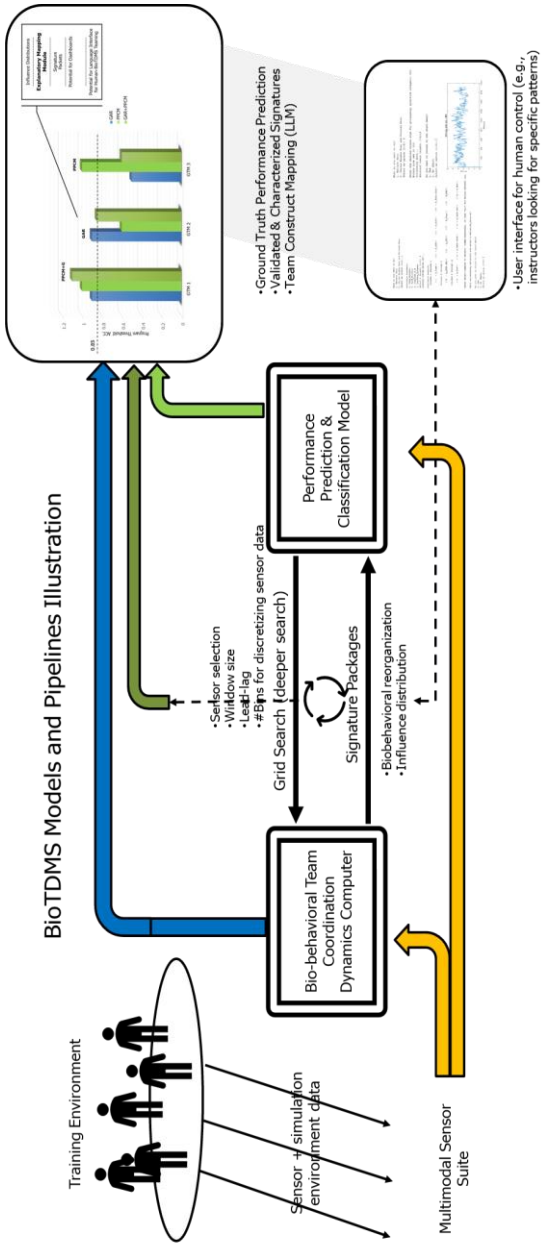


Fig. 1. BioTDMS data processing and modeling pipelines.

Measures

BioTDMS employs an advanced sensor suite to capture physiological, cognitive, and communicative dynamics (Fig. 2). In selecting sensor types, we opted for sensors that are developer friendly to address the need for customization for specific tasks/environments (i.e., the FiST). We prioritized sensors that have been previously demonstrated in studies such as Navy flight simulators with Navy test pilots that minimize donning and doffing time and maximize subject comfort and acceptance. Additional requirements for our sensor suite include the following: system setup plus teardown must be less than 30 minutes; team members should be able to stand up and/or move around. The MMSS has a single data collection laptop for each team member to allow for simultaneous setup, reduce the risk of running sensor software on untested systems, and allow for increased movement.

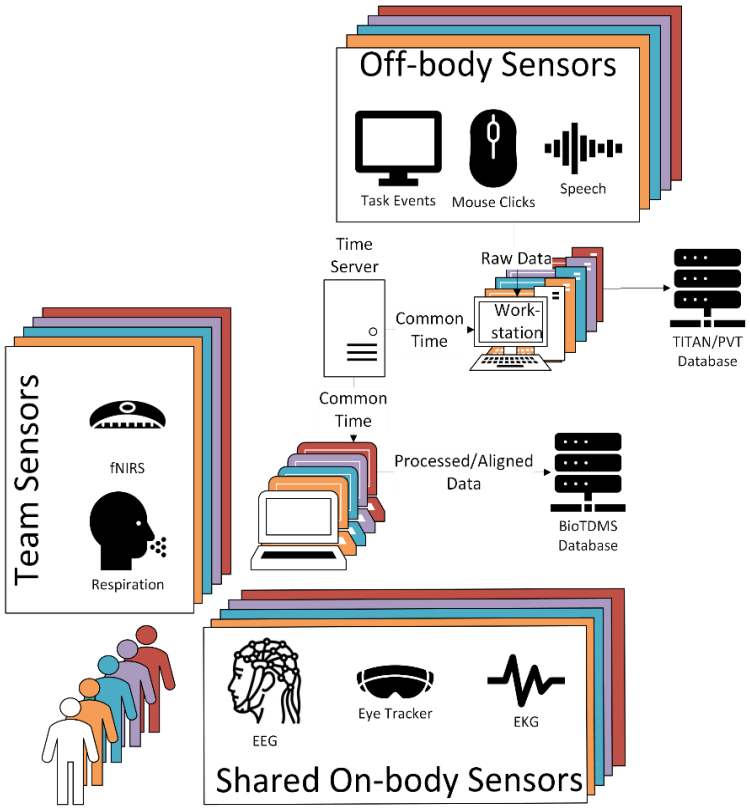


Fig. 2. Multimodal Sensor Suite (MMSS) components.

EEG (Neural Dynamics)

The usage of Advanced Brain Monitoring (ABM) B-Alert EEG system provided robust, artifact-resistant data acquisition under realistic operational conditions. Extensive training enabled the deployment of EEG headsets within 20-30 minutes (other sensors are markedly faster to apply), with validation demonstrating stable signal collection despite substantial participant movement. ABM B-Alert was also chosen to maximize comfort, with the trainees donning the caps at the start of the day and leaving them on for up to 8 hours, during which they participated for approximately 2.5 hours of data collection time.

Cardio-respiratory Dynamics

Using Biosignals Plux units, the system concurrently collects respiration, EKG, and fNIRS data. Additionally, data collected through five ABM EEG units enhance robustness against motion artifacts through redundant data streams, enabling advanced cardiac signal processing and artifact correction strategies. The Plux sensors are quick to apply and are fitted on either side of the forehead held in place by the ABM EEG cap.

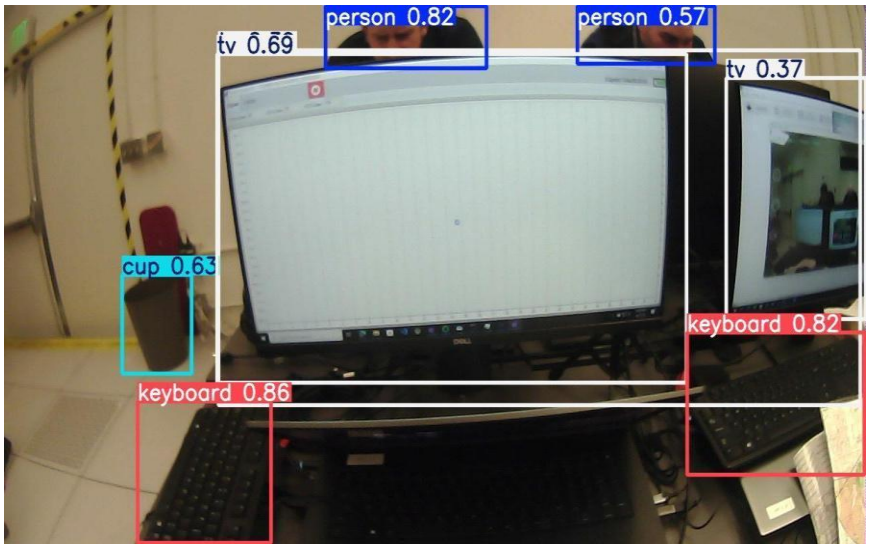


Fig. 3. Example of detected images, including their names along with confidence scores, using the world camera and YOLO model on MMSS eye tracking data.

Eye Tracking (Behavioral Dynamics)

Pupil Labs Core devices capture eye movement, gaze fixation, pupil dilation, and blink rate, providing a window into cognitive load and visual attention. Coupled with object detection using the YOLOv11 model, this data

yields rich behavioral metrics indicating task engagement and attentional allocation. Figure 3 shows image detection using this setup.

Speech (Communicative Dynamics)

For the communication data, we adopt a two-stage, fully automated pipeline that converts raw multi-speaker audio into a time-aligned, speaker-attributed transcript suitable for interaction analysis. For lexical alignment: Audio is first transcribed with OpenAI WhisperX large-v3. WhisperX produces initial sentence-level segments and subsequently forces sub-word alignments, yielding word and sentence level timestamps.

For speaker diarization: The aligned waveform is passed to pyannote. A segmentation network (segmentation-3.0) identifies speaker-change boundaries, after which an embedding network converts each single-speaker segment into a 256-dimensional representation of vocal timbre. Embeddings are clustered by agglomerative/HDBSCAN clustering with the number of clusters constrained to $k = 5$, matching the known participant count and preventing over-segmentation.

For fusion: A lightweight post-processor merges pyannote's speaker turns with WhisperX transcript timings: for every diarization segment all words whose start time $\in$ [start,end) are concatenated. The result is a tabular corpus (Person | Start | End | Text) that preserves the complete dialogue without manual intervention. This stage of the pipeline can also be done completely manually.

Data Collection and Signal Processing

The initial Data Collection Event (DCE1) involved extensive fieldwork at Twentynine Palms, involving four five-member FiST teams over six distinct operational scenarios. Two more data collection events (DCE2, DCE3) are planned for Twentynine Palms. This paper reports findings from DCE1.

Data Acquisition. Native sensor acquisition software provided stability and reliability, with remote monitoring via RealVNC ensuring continuous oversight of data integrity. The collected data was then systematically organized in a structured repository to facilitate offline processing using the MMSS signal processing pipeline.

Signal Processing Pipeline. Rigorous preprocessing pipelines were implemented to enhance data quality. EEG signals are processed using bandpass filtering, notch filtering, and SSP algorithms for artifact removal (eye movements and cardiac artifacts). Power spectral density analyses ensured identification and removal of noise components. EKG and Respiration undergo signal quality assessments to identify periods of unusable data, with alternative estimation methods (respiratory sinus arrhythmia from EKG signals) explored to recover respiratory metrics. fNIRS data undergo spectral analyses to identify reliable versus unreliable signals, isolating valid data for modeling. Eye tracking data are processed using Pupil Labs software and YOLO-based image processing, gaze data were systematically denoised and mapped to relevant environmental objects, facilitating high-quality behavioral analysis. As depicted in Fig. 1, data from the MMSS signal processing pipeline feeds directly into the B-TCDC and PPCM for dynamical systems modeling and performance prediction.

Biobehavioral-Team Coordination Dynamics Computer (B-TCDC)

A central innovation of BioTDMS is the Bio-behavioral Team Coordination Dynamics Computer (B-TCDC), built upon a layered symbolic dynamics modeling framework (Gorman et al., 2019). The primary aim of the B-TCDC is to quantify and interpret complex team dynamics across multiple bio-behavioral modalities by leveraging the concept of "reorganization entropy," a computational metric of adaptation within and across individuals and the team, and "influence." a computational metric of the degree to which team members can change (or be changed by) patterns at the team level. Taken together, the B-TCDC modeling and metrics serve to quantify distributions of team member influence across pockets of adaptation detected from the MMSS data streams in real time.

The symbolic dynamics framework transforms continuous physiological and behavioral data streams (EEG, EKG, respiration, eye-tracking, communication) into discrete symbolic states, facilitating modeling of interactive states (Gorman et al., 2025) of the team across biological, behavioral, and bio-behavioral levels of analyses. This approach dramatically reduces data dimensionality while preserving critical information about the underlying interactive state space. Specifically, by employing symbolic discretization techniques, BioTDMS represents subtle physiological and behavioral changes as symbol sequences, enabling the calculation of entropy and mutual information metrics within a moving-window paradigm.

The layered dynamics model ingests on body (neural activity, EKG, eye tracking, respiration) and off body (speech, task events) sensor data from the MMSS. Each sensor is assigned a variable name q_i. Each q_i can be either a continuous amplitude varying (e.g., EEG) or discretely varying (e.g., speaking or not) sensor. Continuous sensors are discretized into a specified number of bins to code different amplitudes (e.g., low, medium, high amplitude; etc.,) into bins using binary symbols (Fig. 4). Number of bins is an adjustable parameter, where two bins is the simplest coding (i.e., on/off; high/low states) and number of bins equaling the number of possible values the sensor can take being the limit of number of bins (the latter case just reproduces the signal but using discrete binary states). The purpose of binary encoding across sensors is to create mutually exclusive and exhaustive sets, Q ($q_i \subset$ Q), where $\bigcup q_i = Q$ and $q_i \cap q_j = \emptyset$ for all $i \neq j$. Mutual exclusivity of the encoding "alphabet" mathematically ensures computation of unique interactive states across all sensors (Q) at any point in time by "summing" across all sensor states (all q_i) at that time point by horizontally concatenating the q_i states at each time point as illustrated in Fig. 4. Mutual exclusivity also allows the B-TCDC to compute unique interactive states across subsets of sensors (e.g., for one team member or for all bio vs. behavioral sensors) to analyze the dynamics of subsets in the context of the larger interactive space. The interactive state space, used for the results in this paper, utilized all four bins defined by quartiles of the observed signal. For amplitudes, these bins can be interpreted as "very high." "somewhat high," "somewhat low," and "very low," The symbolic interactive states are observed over time, creating symbolic

interactive state time series that are analyzed for reorganization entropy and influence across the entire system and for subsets of the system, including the five FiST members (Lead, FSO, FOA, FOM, JTAC) and different bio-behavioral layers (neural, autonomic, speech).

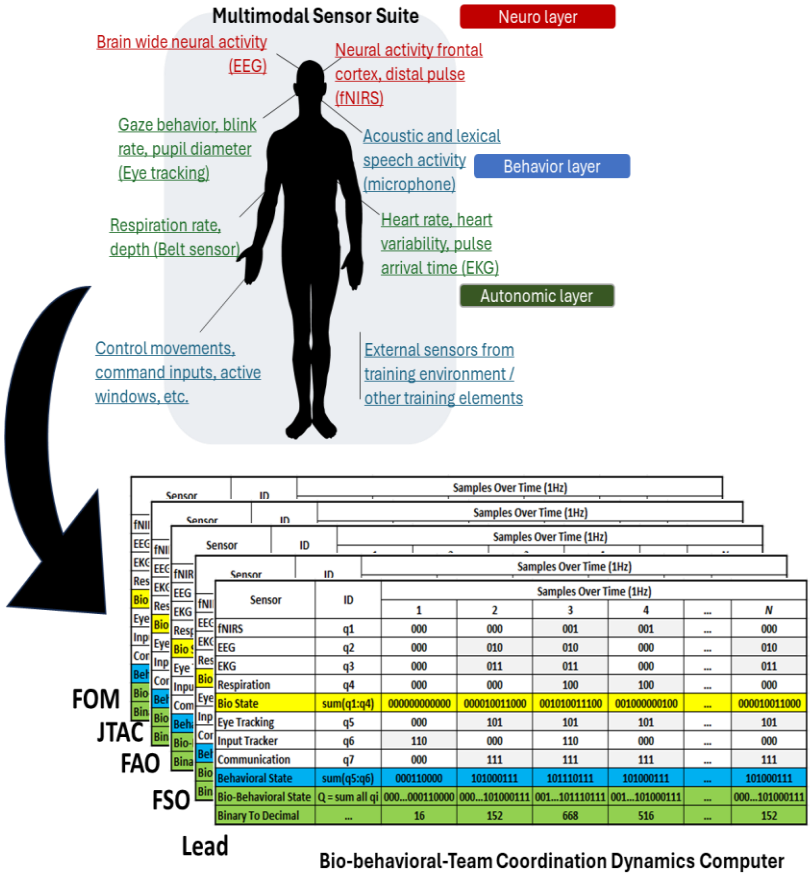


Fig. 4. BioTDMS Interactive State Model illustrating biological and behavioral layers across FiST team members; overall team states are concatenated across all sensors and all team members (not shown). This figure is illustrative and not generated using real data.

Capitalizing on grouping mutually exclusive sensor encodings, the B-TCDC explicitly comprises three interrelated symbolic dynamics layers, each corresponding to a distinct domain of bio-behavioral functioning. The *neural layer* employs EEG and fNIRS data to capture neural signatures indicative of

cognitive adaptability, attention dynamics, workload fluctuations, and task engagement. Artifact-corrected EEG and fNIRS signals undergo symbolic discretization, with entropy-based metrics quantifying changes in neural coordination, particularly in response to task perturbations ("training injects") that necessitate rapid cognitive reorganization.

The *autonomic layer* captures physiological adaptation through integrated cardio-respiratory signals (EKG, respiration, fNIRS). The autonomic layer analyzes heart rate variability (HRV), respiratory synchrony, and blood-oxygenation dynamics to produce symbolic representations that reflect physiological stress, recovery, and adaptation processes at the individual and team levels. Reorganization entropy (described later) within this layer quantifies the team's physiological capacity to respond effectively to external demands and internal coordination challenges.

Currently under final integration, the *speech layer* employs verbal and nonverbal communication data, including speech patterns, interaction frequency, and conversational turn taking (e.g., Gorman et al., 2020). By capturing symbolic communication states, this layer can assess adaptive shifts in team communication strategies, patterns of dominance, collaborative problem-solving behaviors, and overall communication efficacy (e.g., Gorman et al., 2025; Gorman et al., 2020; Reitman et al., 2025).

Each layer independently generates moving window entropy metrics to quantify reorganization (Grimm, Gorman Cooke, Demir, & McNeese, 2023; Stevens et al., 2016) reflecting the amount of continuous adaptation within that layer and the team overall. Reorganization entropy is obtained by computing Shannon entropy across the interactive states in that layer or the team overall (**Q**) in a moving window of size *ws*, where *ws* is an adjustable parameter. Theoretically, reorganization entropy is tied to the *general adaptive response* (GAR; Gorman, Grimm, & Dunbar, 2018; Gorman et al., 2020; see Selye, 1950 for a medical definition; e.g., Fig. 5), which is a response mechanism that predicts survival and success in living systems and systems that interact with living components and represents a team cognitive skill that can be objectively quantified across different team training domains (e.g., surgical teams and SPAN, Gorman et al., 2020; en route CCAT teams, Grimm, Gorman, Robinson, & Winner, 2022; UAV teams, Grimm et al., 2023). GAR is typically measured in response to perturbations, novel training challenges, stressors, and subtask transitions, which make it ideal for measuring team cognitive skill in event-based simulation training environments. However, the true analytical power of the B-TCDC emerges through cross-layer integration, producing comprehensive reorganization entropy metrics for the team overall, each team member, and across different model layers (Fig. 6). These integrated measures holistically represent a team's overall bio-behavioral adaptation, combining cognitive, physiological, and communicative dimensions into singular performance predictors that help trainers assess a team's ability to not only adapt individually but together as a team.

In preliminary analyses using FiST data from Twentynine Palms, reorganization entropy metrics from the Neuro and Autonomic layers correlated strongly with subjective instructor ratings, demonstrating that high entropy values, indicative of greater adaptive flexibility, were consistently associated with superior team performance. These results are highlighted below. Furthermore, ongoing integration of the Communicative layer, in future works, is expected to enrich these predictive relationships substantially.

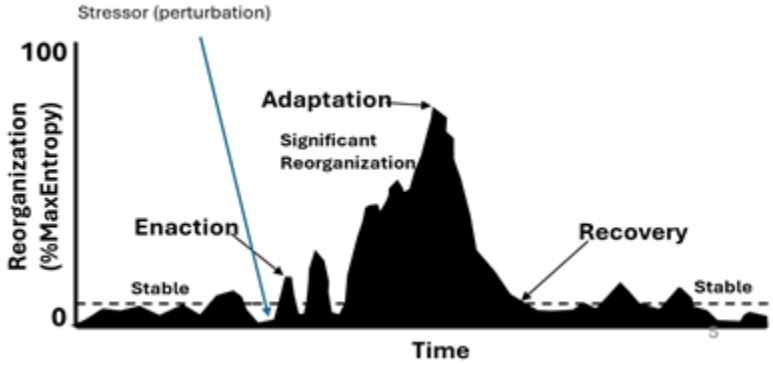


Fig. 5. Theoretical components of the General Adaptive Response (enaction, adaptation, recovery) from bio-behavioral reorganization time series with reference to stressors (perturbations), task/subtask transitions, and event-based training injects.

Beyond reorganization entropy, the B-TCDC also computes "influence distributions," quantifying how individual team members impact adaptation across these layers. By explicitly modeling influence dynamics, the B-TCDC provides actionable insights into critical team roles and interactions that characterize team adaptability, enabling targeted training interventions aimed at optimizing team composition, role assignments, and adaptive capacities.

Performance Prediction and Classification Model (PPCM)

We developed a Performance Prediction and Classification Model (PPCM) that integrates multimodal physiological signals from the BioTDMS framework into a machine learning pipeline to predict team performance outcomes. The goal of this model is to determine the extent to which individual-, interactional-, and team-level physiological indicators can accurately predict team effectiveness and relevant cognitive state measures.

As shown in Fig. 7, we adopted the multimodal physiological and behavioral signals from the Multimodal Sensor Suite (MMSS) for five team roles: Lead, JTAC, FSO, FOM, and FOA. Each team member was instrumented with sensors capturing ECG, EEG, RSP, and eye tracking, alongside synchronized voice communication recordings. Physiological data were down sampled to 1 Hz

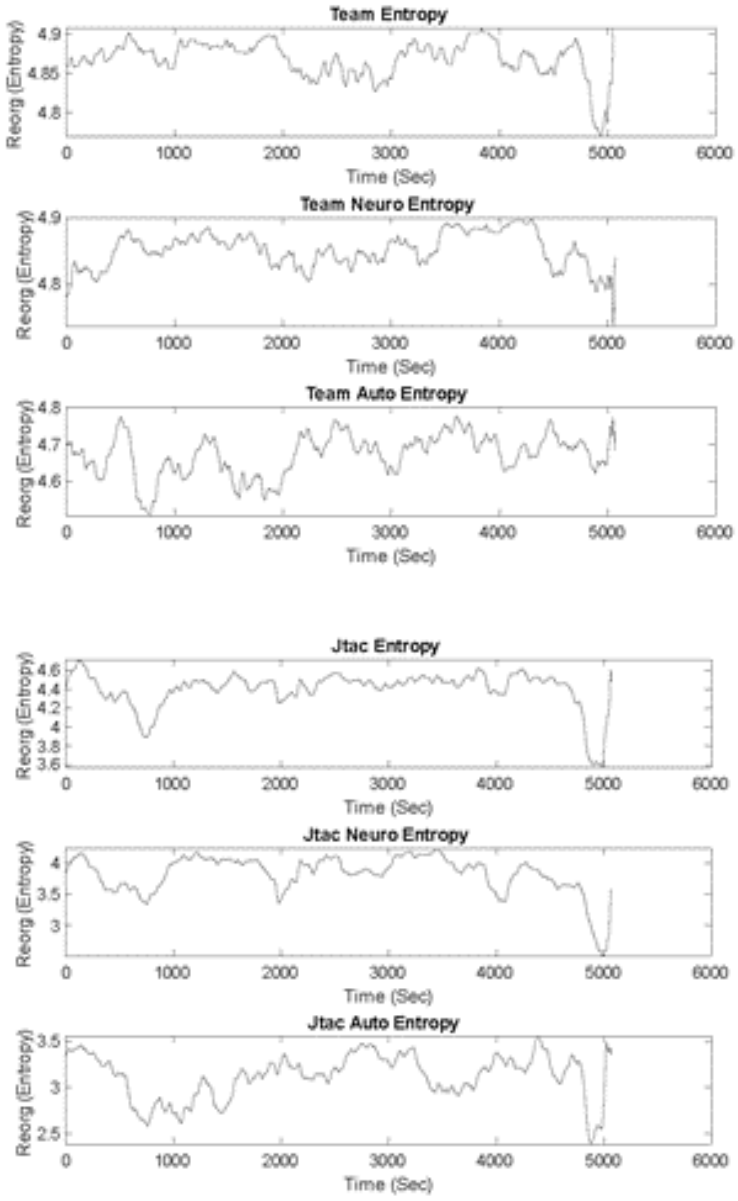


Fig. 6. Example reorganization entropy metrics computed at Neuro, Autonomic (Auto), and Overall for a team one of the team members (JTAC).

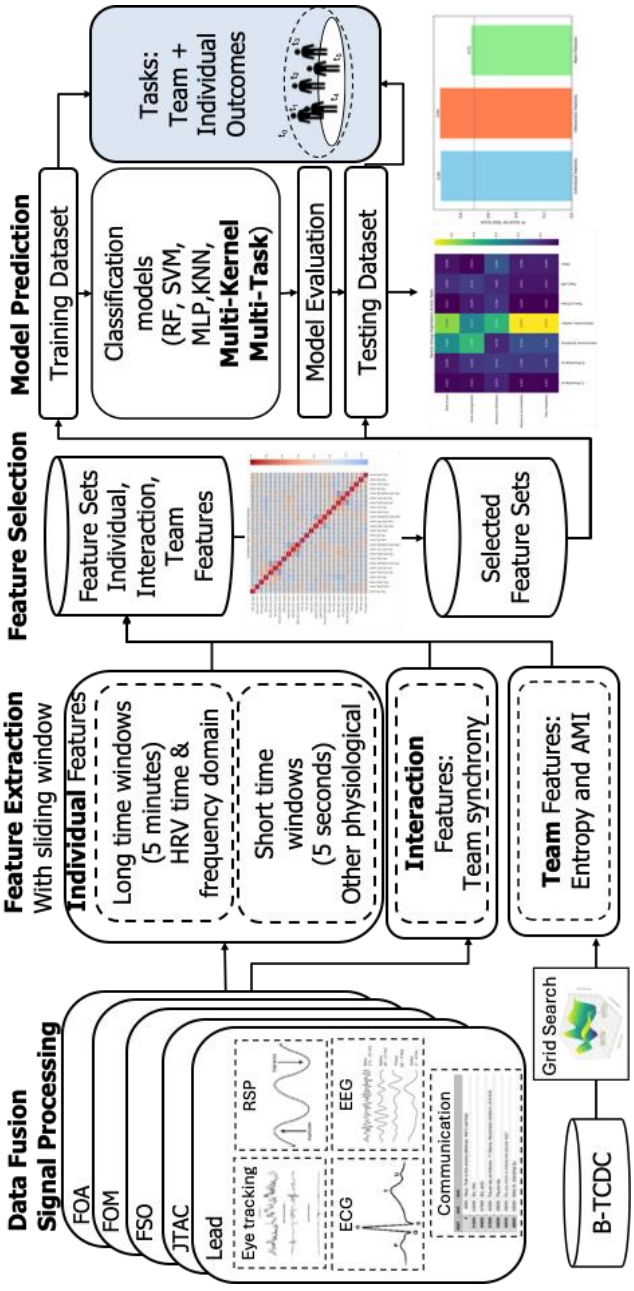


Fig. 7. Performance Prediction and Classification Model (PPCM) Pipeline.

to expedite preprocessing data to validate a feature set. These data streams were processed using both short- and long-window analyses. Short-window (30-second) segments were used to extract dynamic features such as instantaneous heart rate and respiratory variability, while long-window (5-minute) segments were used to compute heart rate variability (HRV) features in both time and frequency domains. Features were categorized at three levels: individual, interactional, and team-level. Individual features were extracted from each modality using time-domain, frequency-domain, and nonlinear analyses.

Interaction features captured physiological synchrony between team members within the same modality (e.g., Lead–JTAC RSP synchrony), computed using sliding-window correlation techniques. Team-level features were derived using the B-TCDC module, which quantified team entropy and average mutual information (AMI) across physiological channels.

A total of 850 candidate features were extracted. A multi-stage feature selection process was then applied via correlation-based pruning to mitigate multicollinearity. This pipeline yielded a reduced, high-utility set of 211 features optimized for predictive modeling. The final feature set was used to train supervised machine learning models aimed at predicting both team-level and individual-level cognitive and performance outcomes. We compared several classification and regression algorithms, including Random Forest (RF), Support Vector Machines (SVM), Multi-Layer Perceptrons (MLP), and k-Nearest Neighbors (KNN). Our primary modeling framework employed a multi-kernel multitask learning approach to jointly predict multiple outcome variables while leveraging shared structure across tasks. We employed a group 5-fold cross validation by utilizing the randomized runs, ensuring it is predicting on 'unseen' runs. Model performance was assessed using metrics R^2 and RSME. Feature importance was evaluated using model-specific importance metrics.

Ontological Mapping Module for Explainability

In addition to the above, to ensure interpretability and explainability of the BioTDMS predictive models, a sophisticated ontology was developed using OWL taxonomy (World Wide Web Consortium, 2012). This ontology provides semantic alignment between observed bio-behavioral signatures and established team performance constructs (e.g., shared cognition, coordination, decision-making). Natural language processing (NLP) techniques (e.g., contextualized construct pepresentation) were piloted to link scenario competencies explicitly to underlying team constructs, thereby grounding model interpretations in established scientific theory and practical relevance.

RESULTS

The initial validation of the Bio-behavioral Team Dynamics Measurement System (BioTDMS) was conducted during a rigorous field-testing event (Data Collection Event 1, DCE1) at the U.S. Marine Corps Air-Ground Combat Center, Twentynine Palms, California. Teams, each comprising five

members (Lead, JTAC, FSO, FOA, and FOM), participated in simulation scenarios, enabling comprehensive testing of BioTDMS capabilities in realistic, high-fidelity operational conditions. We collected data during eight training simulation scenarios overall, comprising data from four different FiSTs. Simulation scenarios lasted approximately 60 minutes.

The data collection employed the multimodal sensor suite, which included EEG, EKG, respiration, fNIRS, and eye-tracking systems, systematically deployed within approximately 30 minutes per team. Despite challenging conditions, including significant participant movement and unanticipated interactions, the sensor systems demonstrated satisfactory stability, yielding high-quality datasets amenable to subsequent detailed analysis.

Sensor performance monitoring via RealVNC revealed qualitative insights into head and eye movement artifacts, which were effectively mitigated through rigorous preprocessing techniques such as artifact removal algorithms and signal source projections (SSP). Notably, while some respiration and fNIRS signals encountered intermittent noise and dropouts, redundancy in multimodal recordings effectively maintained overall data integrity.

Scenario 1

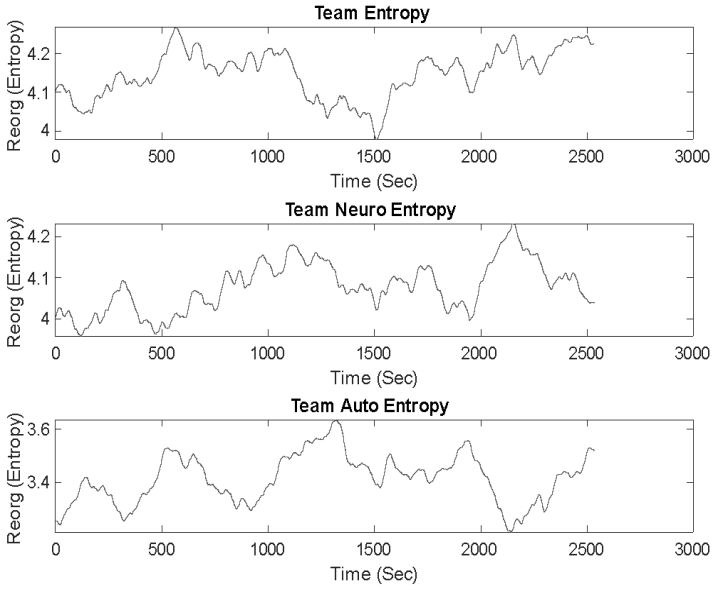
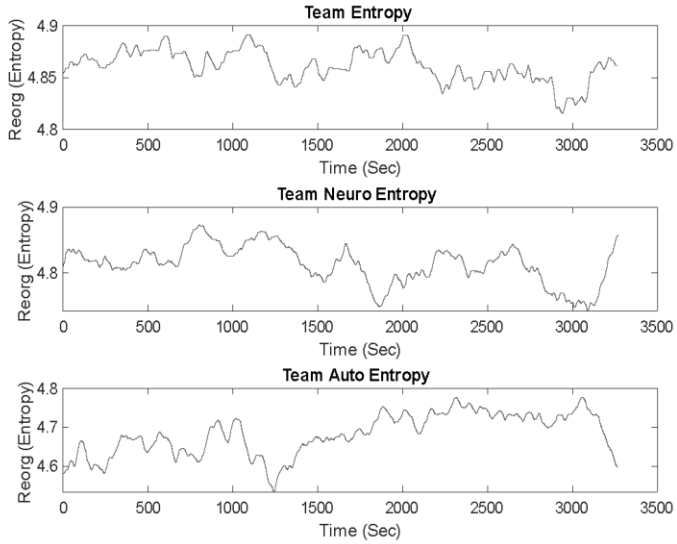


Fig. 8. Plots of team reorganization entropy (higher = more adaptation) across Neuro and Autonomic layers and combined (Overall) for three of the eight DCE1 scenario runs. The scenarios increase in difficulty from Scenario 1 to Scenarios 3 and 4.

Scenario 3



Scenario 4

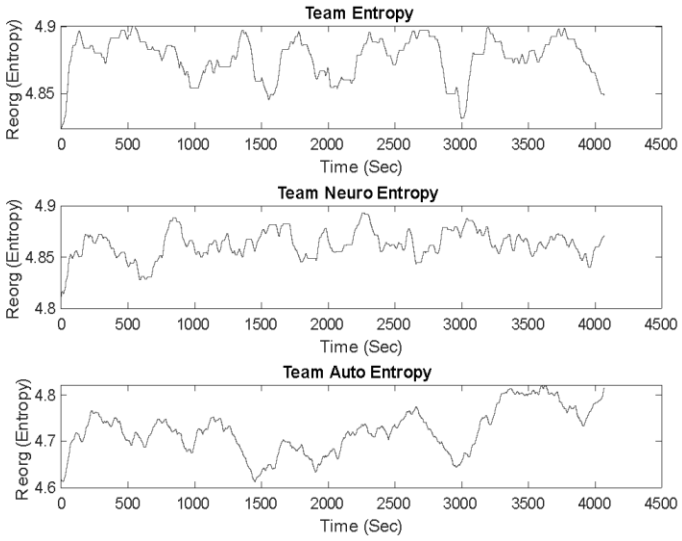


Fig. 8. Continued.

Team Adaptation Metrics: Reorganization Entropy

Central to BioTDMS's analytics pipeline is the Bio-behavioral Team Coordination Dynamics Computer (B-TCDC). Figure 8 shows entropy time series across three types of scenario runs, clearly demonstrating variability in adaptability metrics at both individual and team levels. Specifically, Fig. 8 presents team-level reorganization entropy, highlighting distinctive patterns of adaptation across scenario types. Higher entropy values, indicative of greater adaptability, correlated strongly with improved subjective instructor ratings. These entropy metrics, when correlated with ground truth performance measures, consistently demonstrated strong predictive relationships.

Predictive Analytics and Ground Truth Correlations

Detailed analyses (Figs. 9-11) quantitatively assessed the predictive validity of reorganization entropy. By correlating mean entropy values from Neuro, Autonomic, and Overall layers against established ground truth metrics (GTM) such as instructor ratings (IR), TARGET scores, weighted team scores, and Task Organization Metrics (TOM), the system demonstrated strong predictive power.

As shown in Fig. 9 the overall team entropy explained approximately 50-75% of variance ($r^2 > 0.50$ to $r^2 > 0.75$) across key performance metrics such as weighted team scores and instructor ratings. Notably, the inclusion of holdout data did not significantly alter these correlations, indicating robust model stability.

At an individual level (Figs. 10 and 11), reorganization entropy measures for the JTAC and FiST Lead consistently accounted for significant proportions of performance variance (50-75%), particularly in cognitive and leadership-intensive tasks. The clear differentiation among individual team roles emphasized the sensitivity of BioTDMS to the specific responsibilities and performance dynamics of each role.

These correlations validate that higher reorganization entropy reliably predicts superior team adaptability and thus overall performance, aligning precisely with theoretical expectations regarding the importance of adaptability in team effectiveness.

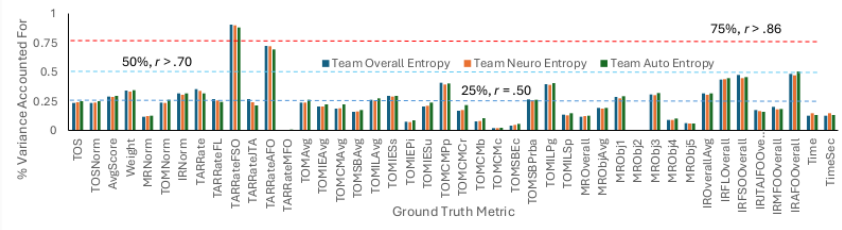


Fig. 9. Mean team reorganization entropy (adaptation): percent variance accounted for, in DCE1 ground truth metrics (GTMs) across neural, autonomic, and combined (overall) B-TCDC layers across n = 8 scenario runs.

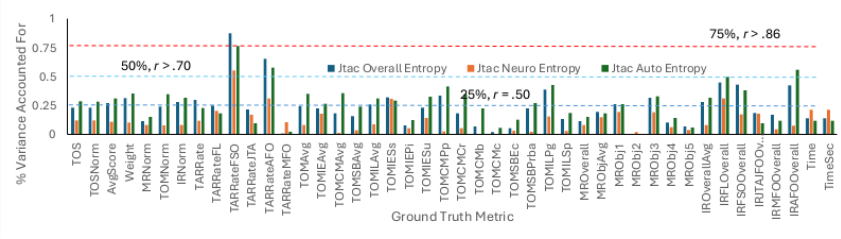


Fig. 10. Mean JTAC reorganization entropy (adaptation): percent variance accounted for, in DCE1 ground truth metrics (GTMs) across neural, autonomic, and combined (overall) B-TCDC layers across *n* = 8 scenario runs.

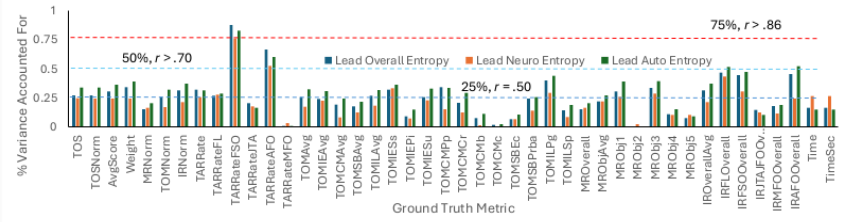


Fig. 11. Mean FiST Lead reorganization entropy (adaptation): percent variance accounted for, in DCE1 ground truth metrics (GTMs) across neural, autonomic, and combined (overall) B-TCDC layers across *n* = 8 scenario runs.

Machine Learning Model Performance

The Performance Prediction and Classification Model (PPCM) pipeline further refined predictive accuracy using multimodal physiological and behavioral data. Specifically, ridge regression models were developed, leveraging a comprehensive feature set comprising approximately 850 initial features (subsequently reduced to 211 highly informative features).

Ridge regression was first trained on individual-level physiological features demonstrating strong performance at a R^2 of 0.88. Incorporating interactional features (e.g., team breathing synchrony, physiological coupling, and cross-correlations) led to a minimal increase to an already strong prediction achieving a R^2 of .90 (see Fig. 12). Additionally, ongoing integration of team features (e.g. entropy & AMI) are being added to enhance the predictive power of the PPCM model.

Feature Importance and Interpretation

A critical analytical outcome was the feature importance analysis (as presented in Fig. 13), which identified specific physiological and interactional metrics that contributed most significantly to model predictions. Ridge Regression's feature importance analysis identified respiratory features as the strongest predictors of team performance.

Figure 13 displays the ten most informative variables ordered by importance where the physiological signals from key team roles, especially leadership positions such as the Lead and JTAC, exert notable influence, with FOA and FOM pupil measures also contributing substantially. Upon doing further analysis, in which we grouped different combinations of physiological sensors, synchrony features notably outperformed most individual-level features when predicting composite measures of cognitive engagement and team coordination. Across the board, we observed that the physiological signals from key team roles, especially those in leadership positions such as the Lead and JTAC, exerted disproportionate influence on overall model accuracy. Notably, the Lead and JTAC sensor groups contribute the most across several critical metrics, with EEG sensors contributing the most to every task. These individuals' physiological states contributed most strongly to the prediction of global team performance and cognitive state indices.

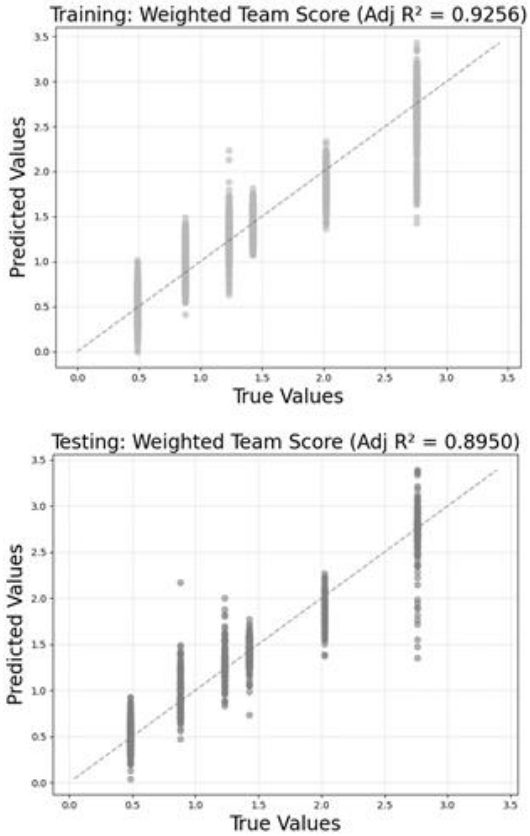


Fig. 12. Upper: Training set results for weighted team score. Lower: Testing set results for weighted team score.

Overall, these findings support the hypothesis that team performance can be effectively modeled using physiological data. While individual metrics are informative, the presence of correlation terms among the top features indicates that coupling within core physiological systems adds predictive value, thereby enhancing the explanatory power of the PPCM. These results lay the foundation for operationalizing BioTDMS data streams into real-time performance monitoring and adaptive team support systems.

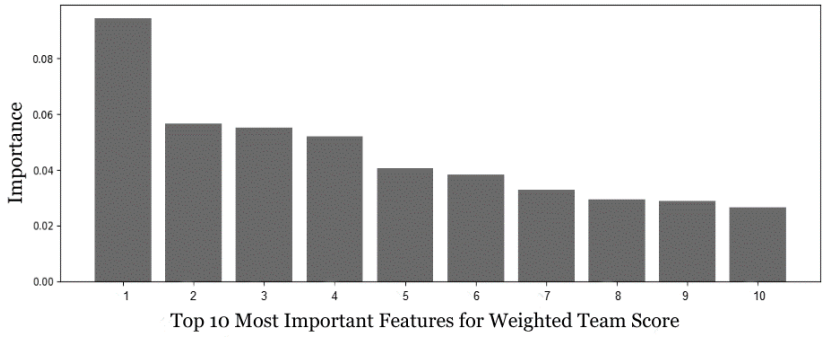


Fig. 13. Feature importance plot from ridge regression for weighted team score. From left to right: (1) Lead Respiration (RSP) Amplitude (mean); (2) Lead RSP RVT (Respiratory Volume per Time, mean); (3) FOA Pupil Diameter (mean); (4) FOM Pupil Diameter (mean); (5) FOA Pupil Diameter (standard deviation); (6) Lead Pupil Diameter (standard deviation); (7) Lead RSP Rate (mean); (8) Lead RSP Rate–RVT Correlation; (9) Lead RSP Rate–Amplitude Correlation; (10) JTAC Mean Inter-Beat Interval (IBI).

Overall, these findings support the hypothesis that team performance can be effectively modeled using physiological data. We predicted that the performance would be further improved when team features are incorporated. While individual metrics are informative, interactional coupling among team members captures additional dimensions of coordination and shared cognitive load, thereby enhancing the explanatory power of the PPCM. These results lay the foundation for operationalizing BioTDMS data streams into real-time performance monitoring and adaptive team support systems.

Summary of Field-Test Findings

In summary, the initial field test results demonstrated BioTDMS's capabilities to quantify complex team performance dynamics that objectively predict ground truth instructor ratings:

1. High-quality multimodal data was successfully collected, processed, and analyzed despite operational challenges.

2. Reorganization entropy emerged as a robust adaptation metric strongly correlated with instructor-rated performance scores, capturing between 50-75% of the variance in team performance across diverse ground truth metrics.

3. Ridge regression analyses further solidified predictive validity, achieving impressive accuracy (adjusted R^2 = .90) by integrating multimodal individual and interactional physiological data.

4. Behavioral data and ontology-based interpretability significantly enhanced the explanatory value of predictions, linking physiological metrics explicitly to cognitive and adaptive team processes.

These results provide compelling validation of BioTDMS as a powerful, reliable, and theoretically grounded tool for objective, real-time assessment of team performance in realistic military training scenarios. We expect to further validate BioTDMS with additional field validations at Twentynine Palms.

The current validation of the BioTDMS demonstrates promising initial results; however, it remains a work in progress with several aspects requiring refinement and further exploration. Notably, variability in predictive accuracy was observed depending on specific sensor combinations and data modalities. For example, physiological data alone provided strong predictive power, yet incorporating interactional features offered incremental improvements. This highlights the complexity inherent in multimodal data integration and suggests opportunities to explore optimal combinations of sensors and signals.

Future directions involve fully integrating the communicative layer, which is expected to substantially enhance the predictive capability and explanatory power of BioTDMS. Overall, continued iterative development and rigorous validation efforts will be crucial in realizing BioTDMS's full potential as a reliable, objective tool for assessing and improving warfighter readiness in dynamic, complex operational contexts

DISCUSSION

This paper has provided a comprehensive overview of the BioTDMS, a state-of-the-art technological solution developed under DARPA's OP TEMPO initiative. By integrating multimodal sensing, symbolic dynamics modeling, and machine learning methodologies, BioTDMS significantly advances the capability to objectively, reliably, and meaningfully assess military team performance.

Linking a sensor suite to a machine learning model through our novel team coordination dynamics computer to optimize performance prediction and classification will revolutionize objective team assessments beyond existing team assessments that rely on subjective observation. While the challenge of collecting bio-behavioral signals has advanced significantly, the challenge of porting them into a reliable, predictive, generalizable team measurement system remains unmet. Most approaches to team measurement aggregate signals over time into static "snapshots" to approximate the team level, rather than measuring team cognition directly (Cooke, Gorman, Myers, & Duran, 2013) and use conventional dynamical models borrowed from other (often nonliving) systems (e.g., synchronization and recurrence) to analyze team dynamics. This is problematic because team dynamics are notably nonstationary and continuously reorganizing, which requires multiscale (individual, team levels; timescales of analysis), multimodal (bio and behavioral), moving window approaches (Gorman &

Wiltshire, 2024) to detect "cross-level" bio-behavioral signatures (e.g., Gorman et al., 2016). BioTDMS, grounded in theories of interactive team cognition (Cooke et al., 2013) and real-time team cognition (Gorman et al., 2020), solves this problem by combining rigorous neural and physiological measurement techniques with team coordination dynamics and machine learning techniques developed for generalizable team cognition and performance assessment (Li, Erickson, Cross, & Lee, 2022).

Initial field tests during Fire Support Team (FiST) exercises at Twentynine Palms yielded compelling preliminary results, confirming BioTDMS's predictive accuracy in accounting for approximately 89% of the variance in instructor-assessed performance scores. Importantly, these results were achieved despite practical challenges, including sensor noise and movement artifacts, underscoring the robustness of our multimodal and integrative approach.

BioTDMS not only demonstrates exceptional predictive potential but also seeks to provide deep explanatory insights into the underlying dynamics of teamwork, adaptability, and team-member influence based in team science through its explanatory mapping module. Through the explicit incorporation of ontology-based mapping, BioTDMS ensures interpretability, linking performance predictions directly to established theoretical constructs within the science of teams, thereby grounding analytics in practical, actionable frameworks.

Future efforts will further enhance the BioTDMS system by fully integrating the communication modality into the symbolic dynamics framework, refining predictive algorithms through additional field data, and deploying real-time analytic capabilities. Moreover, planned expansions into multi-ship aviation operations and human-autonomy teaming contexts will assess BioTDMS's generalizability across diverse operational domains, firmly establishing its relevance as a critical tool for objective team assessments.

BioTDMS is designed to increase the scale of DoD training for the efficient use of training resources to support warfighter readiness and increase training effectiveness by providing instructors with objective metrics to back up their observations and objective feedback for trainees targeting adaptive skills that transfer to the post-training environment. BioTDMS is designed to target generalizable team competencies for broad application across DoD training and tactical domains. In summary, BioTDMS represents a significant advance in how military teams are evaluated, trained, and optimized using theory and analytical techniques of team science, nonlinear dynamics, and machine learning. BioTDMS will provide instructors and decision-makers with precise, objective, real-time insights essential for enhancing warfighter readiness and effectiveness in complex, dynamic military environments.

ACKNOWLEDGMENT

This research was funded by DARPA Contract No. HR001124C0495. The views expressed in this paper are those of the authors, and do not necessarily reflect those of DARPA or the United States Government.

REFERENCES

Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive team cognition. *Cognitive Science, 37*, 255-285.

Gorman, J. C., Amazeen, P. G., & Cooke, N. J. (2010). Team coordination dynamics. *Nonlinear Dynamics, Psychology, and Life Sciences*, *14*(3), 265-289.

Gorman, J. C., Demir, M., Cooke, N. J., & Grimm, D. A. (2019). Evaluating sociotechnical dynamics in a simulated remotely-piloted aircraft system: A layered dynamics approach. *Ergonomics, 62*, 629–643.

Gorman, J. C., Grimm, D. A., & Dunbar, T. A. (2018). Defining and measuring team effectiveness in dynamic environments and implications for team ITS. In J. Johnston, R. Sottilare, & A. M. Sinatra (Eds.), *Building intelligent tutoring systems for teams: What matters* (pp. 55-74). Binley, UK: Emerald Publishing Limited.

Gorman, J. C., Grimm, D. A., Robinson, F. E., Winner, J. L., Wiese, C. W., & Roudebush, C. (2025). Dynamic measures of team adaptation. *Human Factors*, *67*, 123-145.

Gorman, J. C., Grimm, D. A., Stevens, R. H., Galloway, T., Willemsen-Dunlap, A. M., & Halpin, D. J. (2020). Measuring real-time team cognition during team training. *Human Factors*, *62*, 825-860.

Gorman, J. C., Martin, M. J., Dunbar, T. A., Stevens, R. H., Galloway, T. L., Amazeen, P. G., & Likens, A. D. (2016). Cross-level effects between neurophysiology and communication during team training. *Human Factors, 58*, 181-199.

Gorman, J. C., & Wiltshire, T. J. (2024). A typology for the application of team coordination dynamics across increasing levels of dynamic complexity. *Human Factors*, *66*, 5-16.

Grimm, D. A., Gorman, J. C., Cooke, N. J., Demir, M., & McNeese, N. J. (2023). Dynamical measurement of team resilience. *Journal of Cognitive Engineering and Decision Making*, *17*, 351-382.

Grimm, D. A., Gorman, J. C., Robinson, E., & Winner, J. (2022). Measuring adaptive team coordination in an enroute care training scenario. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66,* 50-54.

Li, M., Erickson, I., Cross, V., & Lee, J. D. (2022). Estimating trust in conversational agent with lexical and acoustic features. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 66,* 544-548.

Reitman, J. G., Harrison, J. L., Gorman, J. C., Lieber, R., & D'Mello, S. K. (2025). Communicative influence: A novel measure of team dynamics that integrates team cognition theory with collaborative problem solving assessment. *Journal of Educational Psychology*, *117*, 134-151.

Selye, H. (1950). Stress and the general adaptation syndrome. *British Medical Journal*, *1*(4667), 1383-1392.

Stevens, R., Galloway, T., Gorman, J., Willemsen-Dunlap, A., & Halpin, D. (2016, June). Toward objective measures of team dynamics during healthcare simulation training. In *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, *5*, 50-54.

World Wide Web Consortium. (2012, December 11). *OWL 2 Web Ontology Language document overview* (2nd ed.). https://www.w3.org/TR/owl2-overview/