Evaluating Stress Manipulations in Scenario-Based Command and Control Training: A Multimodal Neural Network Approach

Marcella Hoogeboom
University of Twente
a.m.g.m.hoogeboom@utwente.nl

Jorn-Jan van de Beld
University of Twente
j.j.vandebeld@utwente.nl

ABSTRACT

Command and Control (C2) in crisis response requires teams to coordinate, process information, make decisions, and take appropriate actions under stress. To effectively prepare C2 teams, scenario-based training is a core instructional approach to train and simulate these real-world pressures by manipulating stress. However, it remains unclear if participants actually experience and respond to these scenarios as intended and thus, whether the required capabilities and skills are effectively trained. In contrast to predominantly qualitative evaluations of scenario design, this study adopts a quantitative, multimodal approach using continuous measures of stress and coordination to evaluate and predict three escalating domestic violence arrest training scenarios: low-, medium-, and high-stress. Physiological stress was indexed via heart rate variability, and team coordination behaviors were systematically video-coded. We use recurrent neural network models to compare physiological-only, behavioral-only, and combined inputs to classify the scenario conditions, thereby advancing more evidence-based scenario design and evaluation of C2 training.

Keywords

Command and Control, scenario-based training, stress, multimodal data, neural network analysis.

INTRODUCTION

Command and Control (C2) teams operate in environments characterized by high levels of uncertainty, time pressure, and rapidly changing circumstances. In crisis response domains such as policing, firefighting, and healthcare/acute medical, teams must constantly gather information, interpret and respond to evolving cues, make good decisions, coordinate their actions, and adapt effectively under stress. The demanding conditions under which C2 operate make crisis response an inherently stressful form of teamwork, where stress can influence cognitive functioning and collective coordination processes (Nieuwenhuys & Oudejans, 2017; Liu & Liu, 2018). As coordination failures under stress may influence safety and operational outcomes, developing teams' capacity to operate effectively under pressure is an important aim of C2 training.

To prepare them, C2 skills and capabilities are commonly developed through scenario-based training (SBT) and simulation exercises ('t Hart & Sundelius, 2013). In such training environments, teams engage in realistic, ecologically valid scenarios that provide a controlled environment in which key stressors and task demands can be varied. As such, SBT offers several advantages: it enables repetition without actually threatening the safety of C2 teams and civilians, supports structured debriefing and learning, and allows instructors to adapt the training according to learners' needs. Research on simulation-based learning highlights that well-designed scenarios can foster adaptive expertise, team coordination, and stress regulation by exposing participants to increasingly challenging conditions (e.g., Kleygrewe et al., 2024). An assumption underlying these practices is that manipulating scenario elements, such as threat intensity and aggression level, translates into meaningful differences in experienced stress and coordination demands of the C2 teams.

However, whether scenario design and stress manipulation actually produce the intended responses and effects remains not well-understood. The evaluation of C2 training typically relies on subjective instructor observations or post hoc self-reports from participants. For example, stress is often measured through self-report scales such as visual analogue stress ratings or retrospective anxiety questionnaires (Andersen et al., 2018; Harris et al., 2023). Coordination, on the other hand, is still mostly captured using survey-based perceptions of team processes and performance (e.g., Marks et al., 2001), rather than through fine-grained, temporal behavioral observation. This provides limited insight into how stress actually unfolds or how coordination patterns shift moment by moment during the scenario(s). As a result, it remains unclear whether stress manipulations in SBT actually lead to increases in experienced stress or to distinct stress–coordination dynamics at the team level. Objective, data-based evidence demonstrating whether scenarios produce measurable differences in physiological stress and coordination patterns is largely lacking, despite recent advances in multimodal analytics pipelines in adjacent simulation domains (e.g., Popov et al., 2026). As a result, it is difficult to determine whether scenario designs achieve their intended effects. This limits our ability to validate the SBT design, systematically adjust difficulty levels, and refine interventions in a data-driven manner.

Recent advances in multimodal data collection offer new opportunities to address this challenge. Wearable physiological sensors provide continuous indicators of important physiological measures such as heart rate variability (HRV), offering continuous, objective indices of autonomic regulation associated with stress (Gedam & Paul, 2021; Laborde et al., 2017). In addition, specialized software for behavioral observation and coding frameworks enable fine-grained and systematic assessment of coordination behaviors, including shifts between different forms of coordination (Kolbe et al., 2013, 2014). Yet, despite the availability of these methods, C2-relevant team processes are still frequently studied through a single modality or qualitatively. More broadly, recent multimodal collaboration reviews show that a wide variety of modality combinations and analytic strategies, while stronger methodological integration is still needed (Esterhazy et al., 2025). A combination of these modalities can capture continuous assessment of embodied stress responses and their influence on actual coordination dynamics in scenario-based C2 training. This type of multimodal data can provide new insights into whether scenario manipulations leave detectable changes in synchronized physiological and coordination patterns, moving toward a more comprehensive understanding of C2 crisis response dynamics (Hoogeboom et al., 2018).

The purpose of this study is therefore to examine whether increasingly stressful scenario conditions in police training (low, medium, high stress) can be predicted from multimodal patterns of team functioning. Specifically, we evaluate whether scenario conditions can be classified using physiological data alone, video-coded coordination alone, or their combination, using recurrent neural network models suited for multi-model/temporal data. By reframing scenario validation as a predictive modelling problem, this study contributes to the development of objective, data-driven assessment methods for C2 training and evaluation. Demonstrating that scenario manipulations produce distinguishable multimodal patterns would strengthen evidence-based scenario design, support more precise adaptation and inclusion of stressors in scenarios, and advance research on how stress and coordination co-evolve in high-stakes C2 environments.

BACKGROUND

Scenario-based Training for C2 Teams

Through immersive live or virtual simulations, SBT mirrors operational environments in which participants must integrate tactical skills, communication, situational awareness, and emotional regulation. The effectiveness of SBT rests largely on its functional fidelity, defined as the extent to which scenarios evoke realistic cognitive, emotional, and coordination demands (Maran & Glavin, 2003). In C2 contexts, this means that scenarios must not only look realistic, but also elicit authentic stress responses and team coordination dynamics. A benefit of SBT is therefore its adaptability: scenario elements can be adapted to match participant expertise and learning goals (Plass & Pawar, 2020), often by manipulating exposure to stress and task complexity.

SBT studies that specifically manipulated threat by a threatening opponent, pressure, or ambiguity have demonstrated that scenario features can influence physiological arousal and decision-making processes (Nieuwenhuys & Oudejans, 2011; Nguyen et al., 2021). For example, studies where threat and anxiety in police training were manipulated showed that high-anxiety conditions can initially degrade perceptual-motor performance (e.g., shot accuracy and attentional control), but that training with anxiety can improve officers' ability to maintain task focus and performance under stress, suggesting that scenario-induced stress impacts both physiological arousal and decision strategies (e.g., enhanced goal-directed control following exposure) when stress cues are present.

However, despite promising extant SBT research, a fundamental question remains unanswered. While scenario intensity is frequently adjusted, it remains unclear whether these manipulations in different scenario's systematically produce distinguishable stress–coordination patterns at the team level. Put differently: when stressors are increased in the scenario-designs, do C2 teams actually respond in measurably distinct physiological and coordination dynamics as intended? Without objective multimodal data linking scenario-level manipulations to team-level responses, the validation of SBT design remains largely subjective rather than empirical.

To date, most empirical SBT studies examine single scenarios, providing limited insight into cumulative or comparative effects of increasing complexity in scenario design; and into how systematically escalating stressors across multiple scenarios influence team-level physiological and coordination dynamics.

Physiological Stress in C2 Training

Stress in operational contexts can be defined as a psychophysiological response that emerges when situational demands exceed perceived coping capacity (Nieuwenhuys & Oudejans, 2017). In training under stress, participants are required to maintain performance while exposed to stress-inducing stimuli that reflect actual operational demands (Kelley et al., 2019).

Acute stressors embedded in SBT may include time pressure, ambiguity, threat, noise, crowding, performance pressure, coordination demands, or novelty of the situation (for overviews see: Wollert & Quail, 2018; Nguyen et al., 2021). Research shows that stress can narrow attention, impair working memory, increase heuristic decision-making, and influence motor execution (Nieuwenhuys & Oudejans, 2010, 2011). However, research also found that moderate levels of pressure may enhance attentional engagement and learning (Di Nota & Huhta, 2019).

From a Cognitive Load Theory perspective (Paas et al., 2003), stressors increase intrinsic individual load by adding informational elements, ambiguity, and consequences that must be processed at the same time. In team contexts, this load is distributed among team members (perhaps in a different way), which can have a spill-over effect on other team members (so-called team stress appraisals: Sassenus et al., 2022) and can impact team collaboration and regulation of the situation (Hataaja et al., 2025).

Adaptivity theory (Plass & Pawar, 2020) suggests that instructional adjustments should be based on measurable sources (e.g., stress, performance) and translated into appropriate adaptations, for example, within SBT.

In practice, police SBT relies heavily on instructor-driven adjustments (Cushion, 2022), often guided by subjective impressions rather than objective, time-sensitive data. This creates a gap between the intended scenario design and measurable stress–coordination outcomes. In C2 teams, stress may influence how members exchange information, anticipate each other's actions, or coordinate task execution. Thus, to examine whether the scenarios are experienced as intended, we argue that both patterns of physiological stress and coordination behaviors can serve as complementary indicators of how teams actually respond to manipulated stressors.

Coordination and De-escalation as C2-relevant Behaviors

Coordination in a team can be conceptualized as orchestrating the sequence and timing of interdependent actions (Kozlowski & Ilgen, 2006; Marks et al., 2001). In high-stakes environments such as C2, coordination is not static but unfolds dynamically as team members constantly align their information processing and action under changing time pressures and uncertainty. To capture such coordination processes in high-stakes teams, Kolbe and colleagues developed the Coordination Behaviour in Action Teams (Co-ACT) framework (Kolbe et al., 2013, 2014). Co-ACT conceptualizes coordination as observable, interactional behaviors through which interdependent team members align their actions in real time.

The Co-ACT framework differentiates explicit and implicit coordination, as well as action versus information coordination. Explicit coordination refers to overt behaviors intended to align team members, such as verbal information exchange, directives, clarifications, confirmations, or planning statements. Implicit coordination, in contrast, consists of more anticipatory adjustments and synchronized behaviors that occur without overt communication, relying on shared mental models and mutual predictability. Information-related coordination involves behaviors aimed at sharing, requesting, clarifying, or evaluating information, thereby supporting situational awareness and threat assessment. Action-related coordination concerns the alignment and sequencing of task execution, including positioning, role division, and mutual adjustment actions.

This distinction is particularly relevant for C2 contexts, where teams must continuously shift between exchanging information (e.g., threat assessment, situational updates) and coordinating action (e.g., positioning, restraining, protective moves). In dynamic crisis situations, coordination is therefore not just about the frequency of communication, but also about how teams balance information processing with timely coordination or synchronized action.

Stressful conditions can impact these coordination patterns. For example, increased threat and time pressure have been associated with more directive and action-focused interaction patterns in high-stakes teams (e.g., Hoogeboom et al., 2018, in medical emergency settings). Increased uncertainty may also stimulate information-seeking behaviors, whereas cognitive overload can impair the anticipatory processes and reduce implicit coordination, as maintaining shared mental models becomes more difficult under strain. From a stress and cognitive load perspective, heightened arousal or stress may narrow attentional focus and shift teams toward more reactive or stimulus-driven, as opposed to goal-driven, coordination dynamics (Nieuwenhuys & Oudejans, 2017).

In addition, within the police context, de-escalation behaviors such as active listening, respectful communication, honesty, and empowerment are important for operational outcomes and perceived legitimacy (Todak & James, 2018). These behaviors capture how officers manage the external interaction with civilians. When stressors increase, the balance between de-escalation, control, and force-related behaviors may shift. As such, these behaviors provide an additional behavioral layer for understanding how stress manipulations result into different CR interactions. Together, the integration of Co-ACT coordination behaviors and de-escalation actions enables a fine-grained, behaviorally anchored examination of how stress influences C2 team dynamics in real time.

Why Predictive Modelling and Multimodality

Multimodal approaches can capture complementary aspects of team functioning and their alignment or divergence over time (Lehmann-Willenbrock & Hung, 2024). Sequential machine learning models (e.g., recurrent neural networks) can learn temporal dependencies and non-linear patterns that are difficult to capture through aggregate statistics alone (Hao et al., 2019). For C2 training evaluation, predictive modelling can serve as an assessment layer: rather than only describing what occurred, models test whether multimodal patterns can predict the different scenario conditions, which may inform scenario design, instrumentation choices, and debriefing.

Research Question

The present study examines whether increasingly stressful scenario conditions in police scenario-based training produce distinguishable multimodal patterns of team functioning. Specifically, we ask:

To what extent can scenario condition (low, medium, and high stress) in police scenario-based training be distinguished on the basis of physiological, behavioral, and combined multimodal team patterns?

To answer this question, we compare the classification performance of models based on physiological data only, behavioral data only, and their multimodal combination using neural network analysis.

METHOD

Design and Participants

Data were collected during two police training days at a specialized training center in the Netherlands. Eleven dyads (22 officers; 95.5% male; *M*age = 37.05, *SD* = 9.16; policing tenure *M* = 13.18 years, *SD* = 8.51) participated. We focus on police dyads in arrest simulations because dyadic teams are a common operational configuration.

Scenarios and Stress Manipulation

Each dyad completed three consecutive scripted domestic-violence arrest simulations with escalating stressors. Scenarios were co-designed with instructors to ensure teamwork demands, operational realism, and controlled escalation of the situation.

Table 1. Scenarios and duration

Scenario	Description	Mean Duration (mm:ss)	SD (mm:ss)
Low Stress (LS)	The male suspect is only verbally aggressive; the spouse interferes from the nearby room	01:57	00:39
Medium Stress (MS)	The male suspect is both verbally and physically aggressive; the spouse actively interferes and disrupts the situation	01:35	00:25
High Stress (HS)	The male suspect is both verbally and physically aggressive, and threatens with a knife; the spouse interferes, disrupts the situation, and throws a beer can	02:10	00:38

Physiological Data (HRV)

Officers wore Medtronic Zephyr BioHarness™ 3.0 chest straps recording ECG (250 Hz). A quiet-standing baseline of 5 minutes was collected before the scenarios. HRV was operationalized as Root Mean Square of Successive Differences (RMSSD), which reflects short-term beat-to-beat variability and is often used as an indicator of parasympathetic regulation. Lower RMSSD values generally indicate reduced variability and relatively higher physiological stress or arousal. RMSSD were extracted in 30 s windows with 1 s step, using Kubios with artifact correction procedures following prior work (Alcántara et al., 2020; Arora et al., 2010). Baseline correction was applied to RMSSD time series to reduce between-person variance.

Behavioral Coding

Multi-camera videos (including chest-mounted GoPro and ceiling-mounted cameras) were synchronized and coded in Observer XT using an extended Co-ACT codebook to capture team coordination (Kolbe et al., 2013, 2014), de-escalation behaviors (Todak & James, 2018), and police actions. All behavioral codes can be found in Table 1, including explicit/implicit action and information coordination (e.g., give instruction, monitoring, gather information), de-escalation behaviors (e.g., listen, honesty, empower), and police actions (e.g., verbal force, pepper spray).

Table 2. Extended Codebook with Corresponding Definitions and Examples

Category	Code	Definition	Example
		Coordination (intra-team)	
Explicit action coordination	Give instruction	Includes directives, commands, or assignments of subtasks	"Add the other arm"
	Planning	Includes the verbalisations of non-immediate considerations regarding what should be done and when, also in the form of questions	"On resistance?"
	Speaking up	Questions and direct remarks regarding procedure and further courses of action, also disagreements, and opinions	"So, we are going to handcuff him?"
Implicit action coordination	Action-related talking to the room	Includes comments on performance of own current behaviour not directed to a specific other team member	"I have control"
	Monitoring	Coded when team members observe the actions of fellow team members and anticipate what they are looking for (but not from the environment → gather information)	Police officer watches fellow team member
	Provide assistance	Includes task-relevant action completed without being asked to do so, backing team members up	Police officer helps fellow team member when handcuffing civilian

Explicit information coordination	Information request	Coded if one directly asks another for (task-relevant) information	"Do you have control?"
	Information evaluation	Statements expressing doubt or assurance regarding the accuracy or source of information	"For the dog handler"
	Information upon request	Coded if one answers a (task-relevant) question asked by a team member	"Here is difficult"
Implicit information coordination	Gather information	Coded when a police officer actively gathers information from the environment (but not from team members → monitoring)	Police officer watches the environment
	Information related talking to the room	Coded when team member appeared to address communication not directed to a specific other team member	"This arm has to go to the back"
	Information without request	Providing information to a team member without being asked to do so	"Knife, knife!"
Other	Closed loop behaviour	Includes nodding, or verbally confirming that the message was understood	"Yes, I saw it"
		De-escalation (outside team)	
De-escalation	Respect	Using a respectful tone towards the civilian	Present unless a police officer showed disrespect
	Listen	Actively listen to the story of the civilian and his view on what happened	"Yes … yes"
	Compromise	Making a compromise with the civilian	-
	Honesty	Includes making promises you can keep, presenting the facts of the case, and telling the consequences of certain actions	"Everything you say can be used against you"
	Empower	Give the civilian different options that he can choose from	"You can either cooperate or we have to use violence"
	Calm	Includes that the police officers regulate their emotions and remain calm	Present unless a police officer got frustrated or angry
	Human	Includes talking to the civilian as your equal, avoiding 'cop talk'	Present unless a police officer used cop talk
	Shoes	Includes emphasising with the civilian by placing yourself in the shoes of the civilian	"I get it"
		Action	
Action from police	Police verbal force	Includes shouting at civilian or bullying	"Drop it! Drop it!"
	Police physical contact	The police officer touches the civilian with no intention to harm or use violence	Police officer physically touches the civilian as preparation towards handcuffing
	Police warning	Includes police officer threatening to use force, and pulling tools	"If you don't cooperate, we have to use force"
	Handcuffs	Police officer handcuffs civilian	Police officer handcuffs civilian
	Pepper spray	Police officer uses pepper spray on civilian	Police officer uses pepper spray on civilian
	Taser	Police officer uses taser on civilian	Police officer uses taser on civilian
	Gun	Police officer uses gun on civilian	-
	Police entering	Includes police officer opening the door towards a room and/or entering it	Police officer opens the front door
	Police physical force (no tools)	Police officer uses physical force on civilian	Police officer has to use force to control the civilian

Action from civilian	Civilian threatening	Includes civilian threatening to use force, and pulling tools	"I will stab you"
	Civilian throwing objects	Civilian uses objects (e.g. beer can) on police officer	Civilian throws cushion towards police officer
	Civilian slamming door	Civilian forcefully closes the door	Civilian forcefully closes the door between the two adjacent rooms
	Civilian shock-knife	Civilian uses shock knives on police officer	-
	Civilian physical force (no tools)	Civilian uses physical force on police officer	Civilian keeps his hands to his chest when police officer attempts to handcuff
	Civilian verbal force	Includes civilian shouting, bullying police officers	"I'm not coming with you!"
	Civilian entering	Includes civilian opening the door towards a room and/or entering it	Civilian opens the door from the second to the first room

Neural Network Modelling

First, for multimodal modelling, physiological and behavioural sequences were standardized to the minimum shared length per team and scenario. Scenario evaluation was performed as a sequence classification problem using Python (PyTorch), predicting the scenario condition (HS/MS/LS) from temporal patterns.

Long Short-Term Memory (LSTM) networks were selected to capture the temporal dependencies in sequential data (Hochreiter & Schmidhuber, 1997). The classification task was formulated as a three-class sequence-classification problem, in which each scenario sequence was labeled as low-stress (LS), medium-stress (MS), or high-stress (HS). The aim of the models was to learn whether temporal patterns in physiology, behavior, or their combination could distinguish among these three scenario conditions. Because the data consisted of temporally ordered observations, LSTM networks were used to model dependencies across successive time steps rather than treating the observations as independent.

Three models were designed:

1. Physiology-only: RMSSD-based HRV features with an LSTM layer followed by a fully connected dense layer.
2. Behavior-only: minutely and systematically video-coded behavioral indicators embedded with a fully connected layer with layer normalization and an LSTM layer followed by a fully connected dense layer.
3. Multimodal: concatenated physiology + embedded behavior features aligned per time step, with an LSTM layer followed by a fully connected dense layer.

These models were trained using cross-entropy loss for unnormalized logits and the AdamW optimiser, because this improves training stability by decoupling weight decay from gradient updates (Loshchilov & Hutter, 2019).

We applied the leave-one-team-out cross-validation across the 11 dyads (Hastie et al., 2009). In each split, all sequences from one team were held out for testing while the remaining teams were used for training.

During evaluation, a softmax layer transformed model outputs to probabilities. Model performance was assessed using timestep-level classification accuracy, that is, the proportion of correctly classified time steps within each held-out scenario sequence. We report both scenario-specific mean accuracy across teams and team-level mean accuracy across scenarios. As the classification involved three scenario classes (LS, MS, HS), chance performance was 0.33. This metric was chosen to preserve the sequential nature of the data, although future work should complement it with sequence-level evaluation and independent hold-out validation. The full script is available upon reasonable request.

Model hyperparameters were empirically optimized for the average test set accuracy across splits. These settings should therefore be interpreted as data-specific rather than universally applicable. In other words, when applying the same modeling approach to a new dataset with different sample size, sequence length, measurement properties, or class balance, hyperparameters would need to be re-considered and likely re-optimized. Specifically, the following parameters were optimized: training epochs, learning rate, LSTM layer size, LSTM layer depth, and embedding size of behavioral indicators.

It is important to note that optimization on average test set performance could lead to overestimation of model performance. Future research with larger samples should therefore include an independent hold-out test set to obtain a less biased estimate of generalization performance.

RESULTS

Table 2 shows the results of the leave-one-team-out cross-validated scenario classification analyses across physiological-only, behavior-only, and multimodal models. The reported values are mean timestep-level classification accuracies (M) and their standard deviations (SD) across teams. The findings reveal differences in discriminability across modalities and classification of the scenario. First, for the high-stress scenario, physiology-only (M = 0.57 ± 0.43) and multimodal models (M = 0.52 ± 0.39) perform above chance (0.33), whereas behavior-only performs poorly (M = 0.04 ± 0.01), suggesting that elevated stress is primarily reflected in physiological patterns rather than coordination behaviors alone. Secondly, for the low-stress scenario, behavior-only (M = 1.00 ± 0.00) shows perfect classification accuracy. Multimodal models (M = 0.52 ± 0.42) show strong classification accuracy, indicating that coordination and de-escalation behaviors are particularly distinctive under low-stress conditions. Third, for the medium-stress scenario, behavior-only performs well (M = 0.72 ± 0.46), while physiology-only (M = 0.07 ± 0.10) and multimodal (M = 0.16 ± 0.27) perform below chance, suggesting that medium stress may not produce a consistent physiological pattern/signature and may represent a less clearly differentiated condition.

Table 3. Leave-One-Team-Out Scenario Classification Accuracy (Mean ± SD) by Modality and Stress Condition

Scenario	Physiology-only	Behavior-only	Multimodal
High-Stress	**0.57 +- 0.43**	0.01 +- 0.01	**0.52 +- 0.39**
Low-Stress	0.28 +- 0.34	**1.00 +- 0.00**	**0.52 +- 0.42**
Medium-Stress	0.07 +- 0.10	**0.72 +- 0.46**	0.16 +- 0.27

Table 4. Leave-One-Team-Out Scenario Classification Accuracy (Mean ± SD) by Modality and Stress for each Team

Team	Physiology-only	Behavior-only	Multimodal
Team 1	0.25 +- 0.41	0.34 +- 0.57	0.27 +- 0.37
Team 10	0.32 +- 0.35	0.67 +- 0.58	0.39 +- 0.50
Team 11	0.38 +- 0.54	0.67 +- 0.58	0.21 +- 0.28
Team 2	0.20 +- 0.34	0.34 +- 0.57	0.35 +- 0.56
Team 3	0.33 +- 0.58	0.67 +- 0.58	0.35 +- 0.56
Team 4	0.30 +- 0.53	0.67 +- 0.57	0.38 +- 0.54
Team 5	0.24 +- 0.23	0.66 +- 0.57	0.81 +- 0.15
Team 6	0.39 +- 0.45	0.67 +- 0.57	0.37 +- 0.26
Team 7	0.41 +- 0.48	0.33 +- 0.58	0.51 +- 0.50
Team 8	0.33 +- 0.58	0.67 +- 0.58	0.33 +- 0.58
Team 9	0.22 +- 0.24	0.67 +- 0.58	0.39 +- 0.21

These findings indicate that stress levels manifest differently across modalities, and that combining modalities does not uniformly improve prediction but rather enhances performance, particularly in high-stress scenarios.

Figures 1, 2, and 3 visualize the model predictions over time for each scenario and team. Using leave-one-team-out validation, the multimodal LSTM achieved above-chance scenario classification for a substantial subset of teams. Mean accuracy (averaged across scenarios within each team) varied across teams, ranging from 0.29 to 0.57. Across all 11 teams, the average of the team means was 0.43, exceeding the chance level of 0.33 for three-class classification.

Team-level results showed that we do not see similar predictions across the scenario's for all teams. For example, several teams achieved moderate-to-high accuracy (e.g., Team 5: 0.57 ± 0.30; Team 3: 0.48 ± 0.17), whereas others were near or below chance (e.g., Team 1: 0.29 ± 0.05; Team 2: 0.30 ± 0.08). This variability suggests that, for some teams, combined physiological–behavioural patterns contain consistent information that differentiates low-, medium-, and high-stress scenarios, whereas for others, scenario-specific signatures are weaker or less consistent. Importantly, beyond the overall accuracy, the temporal graphs themselves provide additional insights. For example, examining when and how predictions shift within a scenario (e.g., a sudden shift to high-stress probability during a low-stress scenario) may reveal critical moments in which coordination or arousal patterns change. Such within-scenario fluctuations can inform a more fine-grained understanding of how stress manipulations unfold dynamically and whether specific events or stressors trigger measurable stress- or behavioral effects.

Future Explorations

As this is a work-in-progress, final model specifications (architecture depth, regularization, and hyperparameters) will be reported in the final version of the paper following completion of systematic tuning and validation.

To finalize our results, we also want to explore why we see certain cut-off points in the scenarios. This helps us to understand the specific behaviors or stressors that influence the experienced stress. Other literature also suggested that, especially for longer scenarios, different phases might exist. For example, 'beginning', 'escalation' and 'peak' (Moreno et al., 2024) or the well-known action vs. transition phases of Marks (Marks et al., 2001). This implies that prediction at the scenario-level might not be accurate enough. Instead, our results may provide insight into phase-specific dynamics within crisis situations, helping to clarify when and how stress–coordination patterns shift over time.

In addition, a methodological consideration concerns the potential for overestimating the models' performance when models are evaluated based on average cross-validated test results. An independent hold-out test set would provide a more unbiased estimate. However, given the limited number of teams (N = 11), holding out additional teams for final testing would have resulted in a very small and probably unrepresentative sample. Future research with a larger sample should include independent test cohorts to strengthen external validity.

CONCLUSION

Based on the analyses provided here, the multimodal combination can predict scenario conditions above chance, indicating that stress manipulations across scenarios can leave detectable patterns when physiological and behavioral coordination/de-escalation processes are analyzed jointly, especially for the low-and high-stress scenarios. To enhance C2 training, our results and multimodal patterns can help to uncover whether stress manipulations actually result in distinct coordination patterns, identify which teams are under- or over-challenged, and adjust stressors or task complexity accordingly in SBT when needed.

REFERENCES

Alcántara, J. M., Camacho-Cardenosa, A., Morán-Camacho, E., Marín-Pagán, C., Muñoz, D., & Garatachea, N. (2020). Heart rate variability in elite police officers: An investigation in real-life and simulated tasks. *Applied Ergonomics, 82*, 102962. https://doi.org/10.1016/j.apergo.2019.102962

Arora, S., Sevdalis, N., Nestel, D., Woloshynowych, M., Darzi, A., & Kneebone, R. (2010). The impact of stress on surgical performance: A systematic review of the literature. *Surgery, 147*(3), 318–330.e6. https://doi.org/10.1016/j.surg.2009.10.007

Cushion, C. J. (2022). Scenario-based training in police education: Pedagogical challenges and instructor decision-making in dynamic learning environments. *Policing: A Journal of Policy and Practice, 16*(4), 709–724.

Di Nota, P. M., & Huhta, J.-M. (2019). Complex motor learning and police training: Applied, cognitive, and clinical perspectives. *Frontiers in Psychology, 10*, 1797. https://doi.org/10.3389/fpsyg.2019.01797

Esterhazy, R., Schneider, B., Cukurova, M., et al. (2025). Advancing Multimodal Collaboration Analytics: A Scoping Review. *Journal of Learning Analytics,* 12(2), 105-124.

Hoogeboom, A. M. G. M., Endedijk, M. Groenier, M., de Laat, S., and van Sas, J. (2018). Using sensor technology to capture the structure and content of team interactions in medical emergency teams during stressful moments. *Frontline Learning Research, 6,* 123–147. https:/doi: 10.14786/flr.v6i3.353.

Gedam, S., & Paul, S. (2021). A review on mental stress detection using wearable sensors and machine learning techniques. *IEEE Access, 9*, 84045–84066. https://doi.org/10.1109/ACCESS.2021.3085502

Hao, Z., Liu, M., Wang, Z., & Zhan, W. (2019). Human behavior analysis based on attention mechanism and LSTM neural network. In *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)* (pp. 346–349). https://doi.org/10.1109/ICEIEC.2019.8784479

Harris KM, Gaffey AE, Schwartz JE, Krantz DS, Burg MM. The Perceived Stress Scale as a Measure of Stress: Decomposing Score Variance in Longitudinal Behavioral Medicine Studies. Ann Behav Med. 2023 Sep 13;57(10):846-854. doi: 10.1093/abm/kaad015. PMID: 37084792; PMCID: PMC10498818.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer.

Hataaja, E., Järvelä, S., & Malmberg, J. (2025). Physiological synchrony and collaborative regulation under stress: A multimodal perspective on team learning. *Learning and Instruction, 92*, 101905. https://doi.org/10.1016/j.learninstruc.2024.101905

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Järvelä, S., Järvenoja, H., & Malmberg, J. (2019). Capturing the dynamic and cyclical nature of regulation: Methodological progress in understanding socially shared regulation in learning. *International Journal of Computer-Supported Collaborative Learning, 14*(4), 425–441.

Kelley DC, Siegel E and Wormwood JB (2019) Understanding Police Performance Under Stress: Insights From the Biopsychosocial Model of Challenge and Threat. Front. Psychol. 10:1800. https://doi.org/10.3389/fpsyg.2019.01800

Kleygrewe, L., Hutter, R. R. C., & Oudejans, R. R. D. (2024). *Stress exposure and performance in virtual reality–based police training: The role of scenario design and physiological responses*. Computers in Human Behavior, 146, 107785. https://doi.org/10.1016/j.chb.2023.107785

Kolbe, M., Burtscher, M. J., & Manser, T. (2013). Co-ACT—A framework for observing coordination behaviour in acute care teams. *BMJ Quality & Safety, 22*(7), 596–605.

Kolbe, M., Grote, G., Waller, M. J., Wacker, J., Grande, B., Burtscher, M. J., & Spahn, D. R. (2014). Monitoring and talking to the room: Autochthonous coordination patterns in team interaction and performance. *Journal of Applied Psychology, 99*(6), 1254–1267. https://doi.org/10.1037/a0037877

Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest, 7*(3), 77–124. https://doi.org/10.1111/j.1529-1006.2006.00030.x

Laborde, S., Mosley, E., & Thayer, J. F. (2017). Heart rate variability and cardiac vagal tone in psychophysiological research: Recommendations for experiment planning, data analysis, and data reporting. *Frontiers in Psychology, 8*, 213. https://doi.org/10.3389/fpsyg.2017.00213

Lehmann-Willenbrock, N., & Hung, H. (2024). A multimodal social signal processing approach to team interactions. *Organizational Research Methods, 27*(3), 477–515. https://doi.org/10.1177/10944281231202741

Lehmann-Willenbrock, N., Hung, H., & Keyton, J. (2017). New frontiers in analyzing dynamic group interactions: Bridging social and computer science. *Small Group Research, 48*(5), 519–531. https://doi.org/10.1177/1046496417718941

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Liu, S., & Liu, Y. (2018). Team stress research: A review and recommendations for future investigations. *Occupational Health Science, 2*(2), 99–125.

Malmberg, J., Järvelä, S., Holappa, J., Haataja, E., Huang, X., & Siipo, A. (2019). Going beyond what is visible: What multichannel data can reveal about interaction in the context of collaborative learning. *Computers in Human Behavior, 96*, 235–245. https://doi.org/10.1016/j.chb.2018.06.030

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. Academy of Management Review, 26(3), 356–376.

Moreno, M., Melo, L. P., Grewal, K., Matin, N., Azher, S., & Harley, J. M. (2024). Analyzing multimodal data to understand medical trainees' regulation strategies and physiological responses in high- fidelity medical simulation scenarios. *Metacognition and Learning, 19*(3), 1161–1213. https://doi.org/10.1007/s11409-024-09403-z

Nieuwenhuys, A., & Oudejans, R. R. D. (2017). Anxiety and performance: Perceptual-motor behavior in high-pressure contexts. *Current Opinion in Psychology, 16*, 28–33. https://doi.org/10.1016/j.copsyc.2017.03.019

Nguyen, Q., Jaspaert, E., Murtinger, M., Schrom-Feiertag, H., Egger-Lampl, S., & Tscheligi, M. (2021). Stress Out: Translating Real-World Stressors into Audio-Visual Stress Cues in VR for Police Training. In C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, & K. Inkpen (Eds.), Human-Computer Interaction – INTERACT 2021 (pp. 551–561). Springer International Publishing. https://doi.org/10.1007/978-3-030-85616-8_32

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1–4. https://doi.org/10.1207/S15326985EP3801_1

Plass, J. L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education*, *52*(3), 275-300.

Salahuddin, L., Cho, J., Jeong, M. G., & Kim, D. (2007). Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 4656–4659). https://doi.org/10.1109/IEMBS.2007.4353378

Sassenus, S., van den Bossche, P., Segers, M., & Kirschner, P. A. (2022). Team stress appraisal: A multilevel perspective on how stress perceptions emerge and spread within teams. *Small Group Research, 53*(6), 829–858. https://doi.org/10.1177/10464964221115038

't Hart, P., & Sundelius, B. (2013). Crisis management revisited: A new agenda for research, training and capacity building within Europe. *Cooperation and Conflict*, *48*(3), 444-461. https://doi.org/10.1177/0010836713485711

Todak, N., & James, L. (2018). A systematic social observation study of police de-escalation tactics. *Police Quarterly, 21*(4), 509–543. https://doi.org/10.1177/1098611118784007

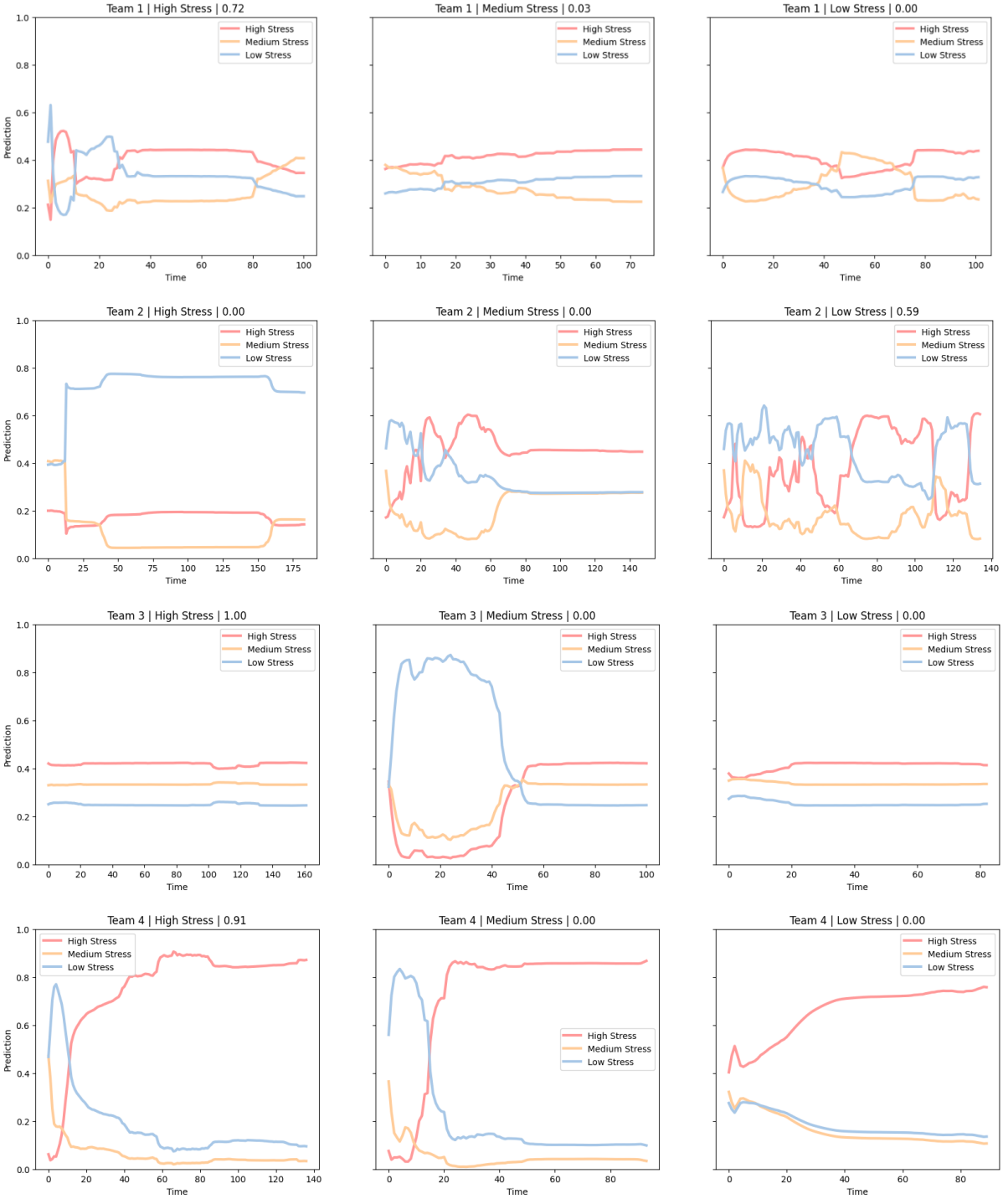
Verhulst, M. J., & Rutkowski, A. F. (2018). Decision-making in the police work force: Affordances explained in practice. *Group Decision and Negotiation, 27*(5), 827–852.

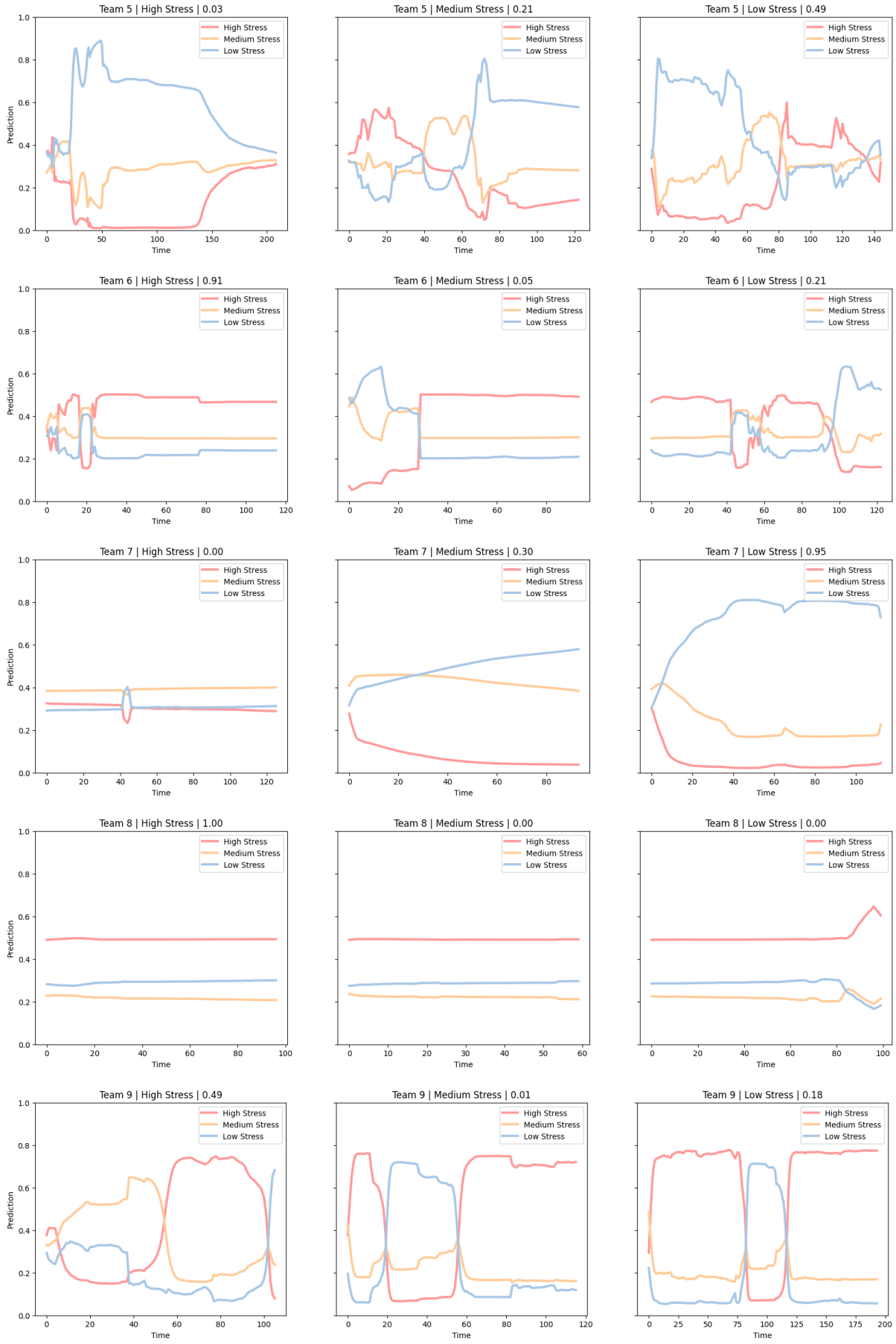
Wollert, T.N. and Quail, J. (2018), A Scientific Approach to Reality Based Training, Three Pistols Publishing, Winnipeg, Manitoba.

Yan, L., Gašević, D., Echeverria, V., et al. (2025). In sync or out of sync? Understanding stress and learning performance in collaborative healthcare simulations through physiological synchrony and arousal. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-025-00475-9

APPENDIX

Figure 1. Timestep-level Scenario Prediction Probabilities using Physiology-only Input for each Team and Scenario




Team 5 | High Stress | 0.03
Team 5 | Medium Stress | 0.21
Team 5 | Low Stress | 0.49
Team 6 | High Stress | 0.91
Team 6 | Medium Stress | 0.05
Team 6 | Low Stress | 0.21
Team 7 | High Stress | 0.00
Team 7 | Medium Stress | 0.30
Team 7 | Low Stress | 0.95
Team 8 | High Stress | 1.00
Team 8 | Medium Stress | 0.00
Team 8 | Low Stress | 0.00
Team 9 | High Stress | 0.49
Team 9 | Medium Stress | 0.01
Team 9 | Low Stress | 0.18
High Stress
Medium Stress
Low Stress
Prediction
Time

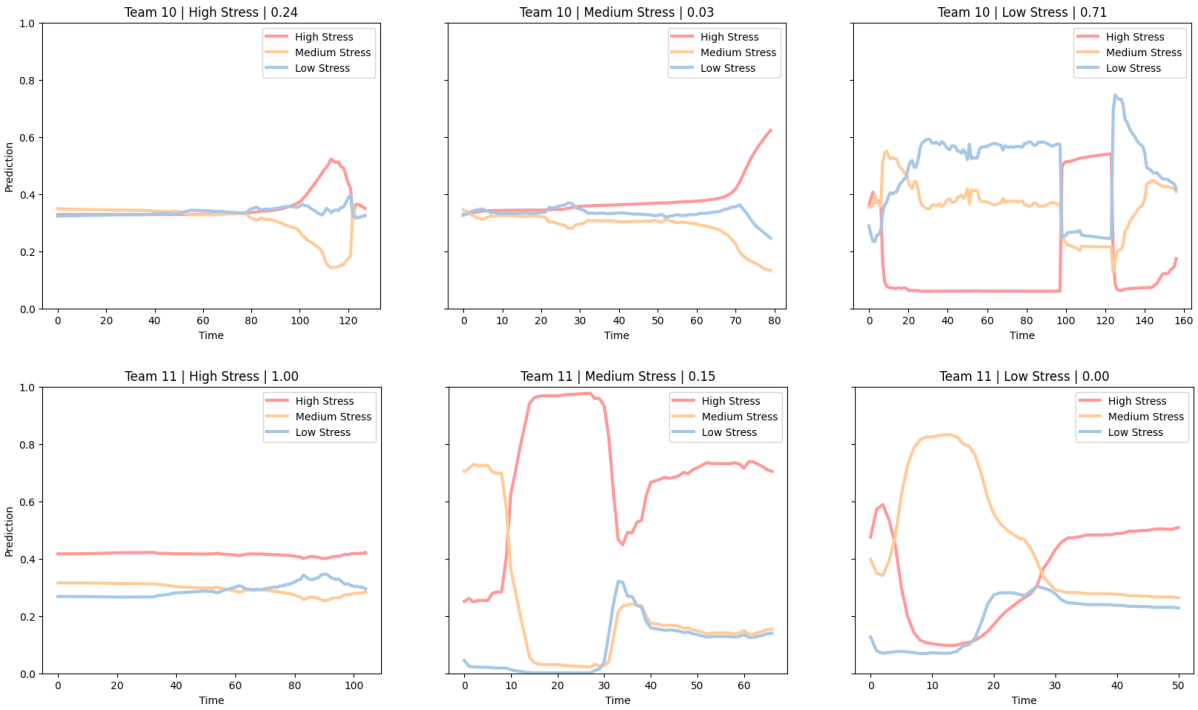
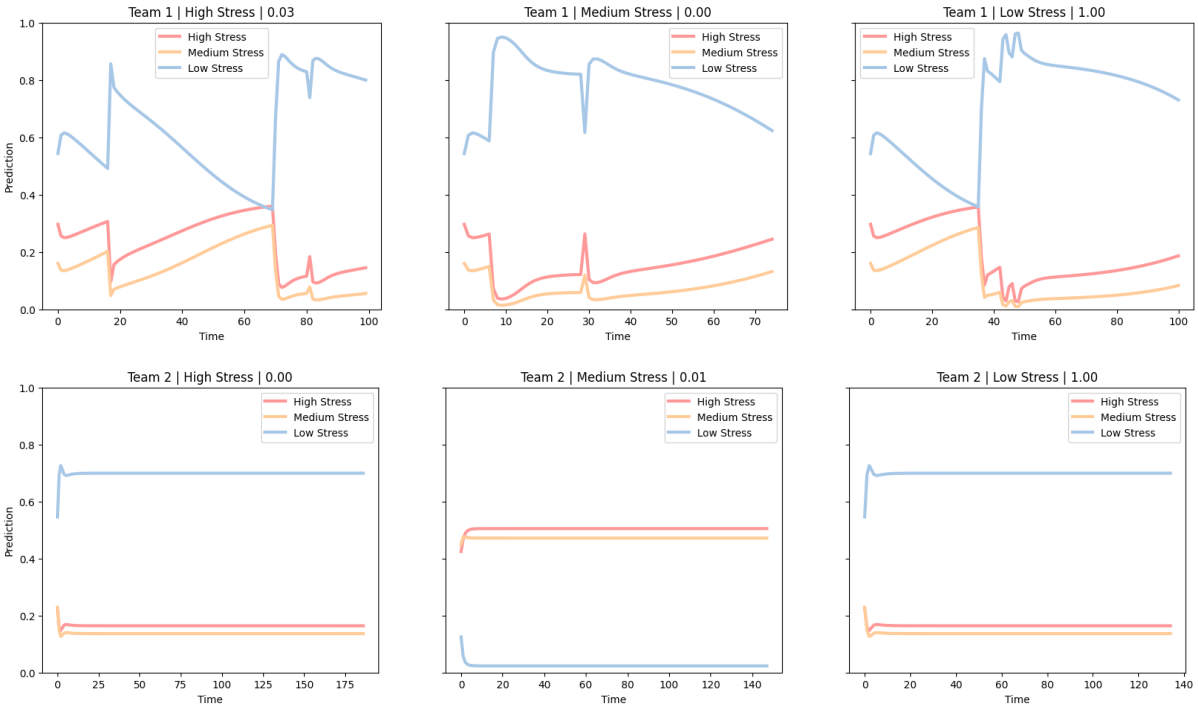
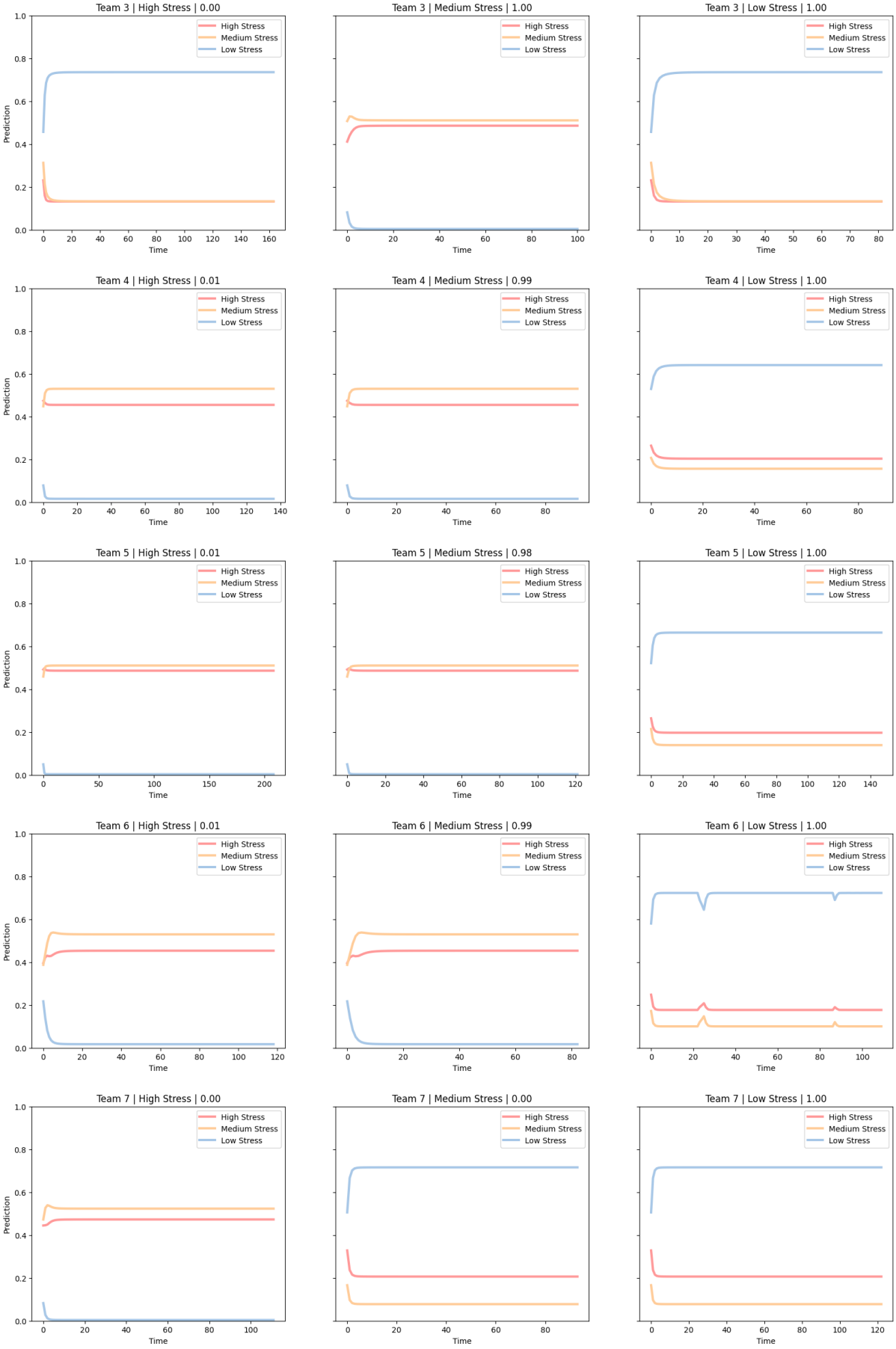


Figure 2. Timestep-level Scenario Prediction Probabilities using Behavior-only Input for each Team and Scenario




Team 3 | High Stress | 0.00
Team 3 | Medium Stress | 1.00
Team 3 | Low Stress | 1.00
Team 4 | High Stress | 0.01
Team 4 | Medium Stress | 0.99
Team 4 | Low Stress | 1.00
Team 5 | High Stress | 0.01
Team 5 | Medium Stress | 0.98
Team 5 | Low Stress | 1.00
Team 6 | High Stress | 0.01
Team 6 | Medium Stress | 0.99
Team 6 | Low Stress | 1.00
Team 7 | High Stress | 0.00
Team 7 | Medium Stress | 0.00
Team 7 | Low Stress | 1.00
High Stress
Medium Stress
Low Stress
Prediction
Time

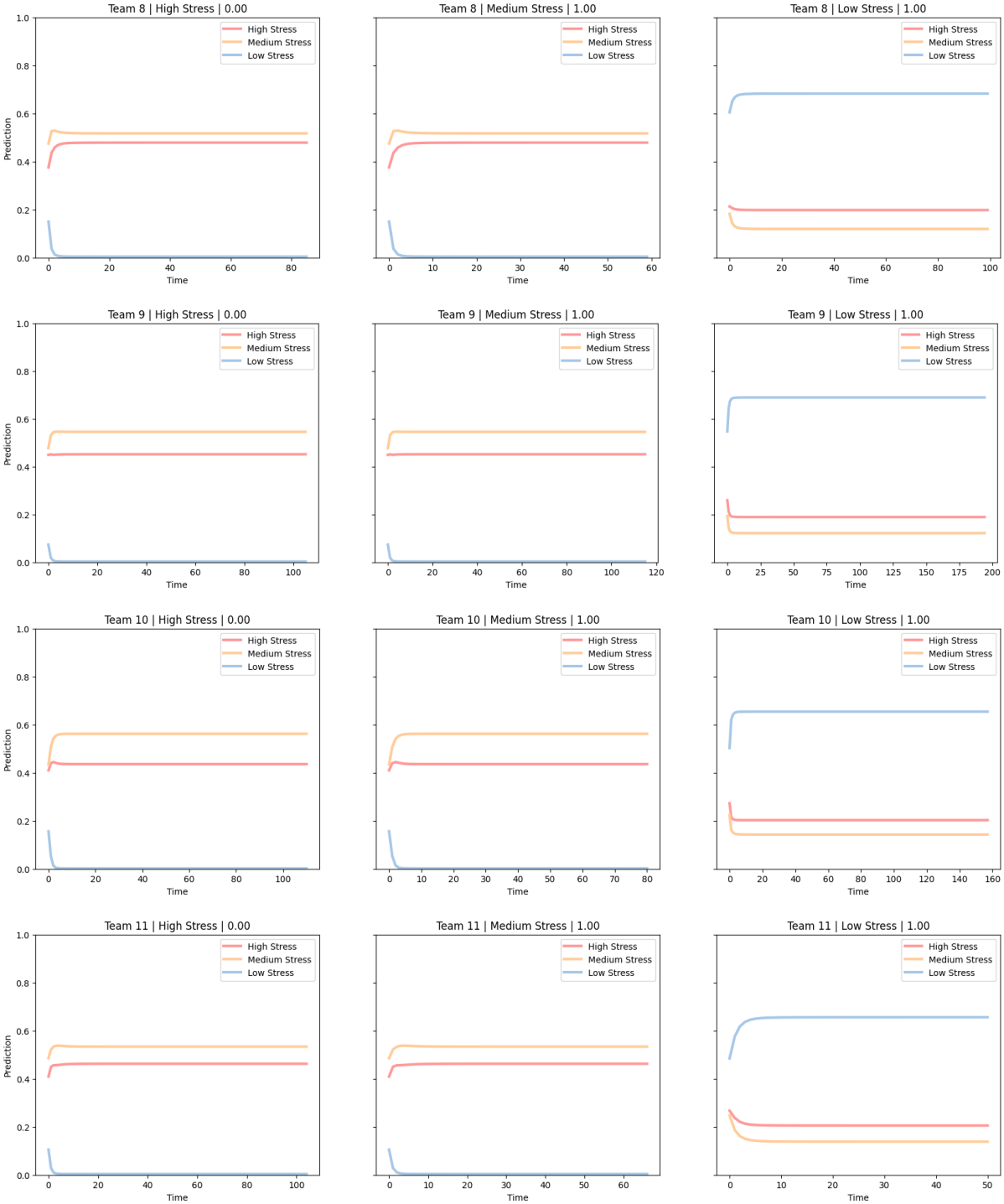
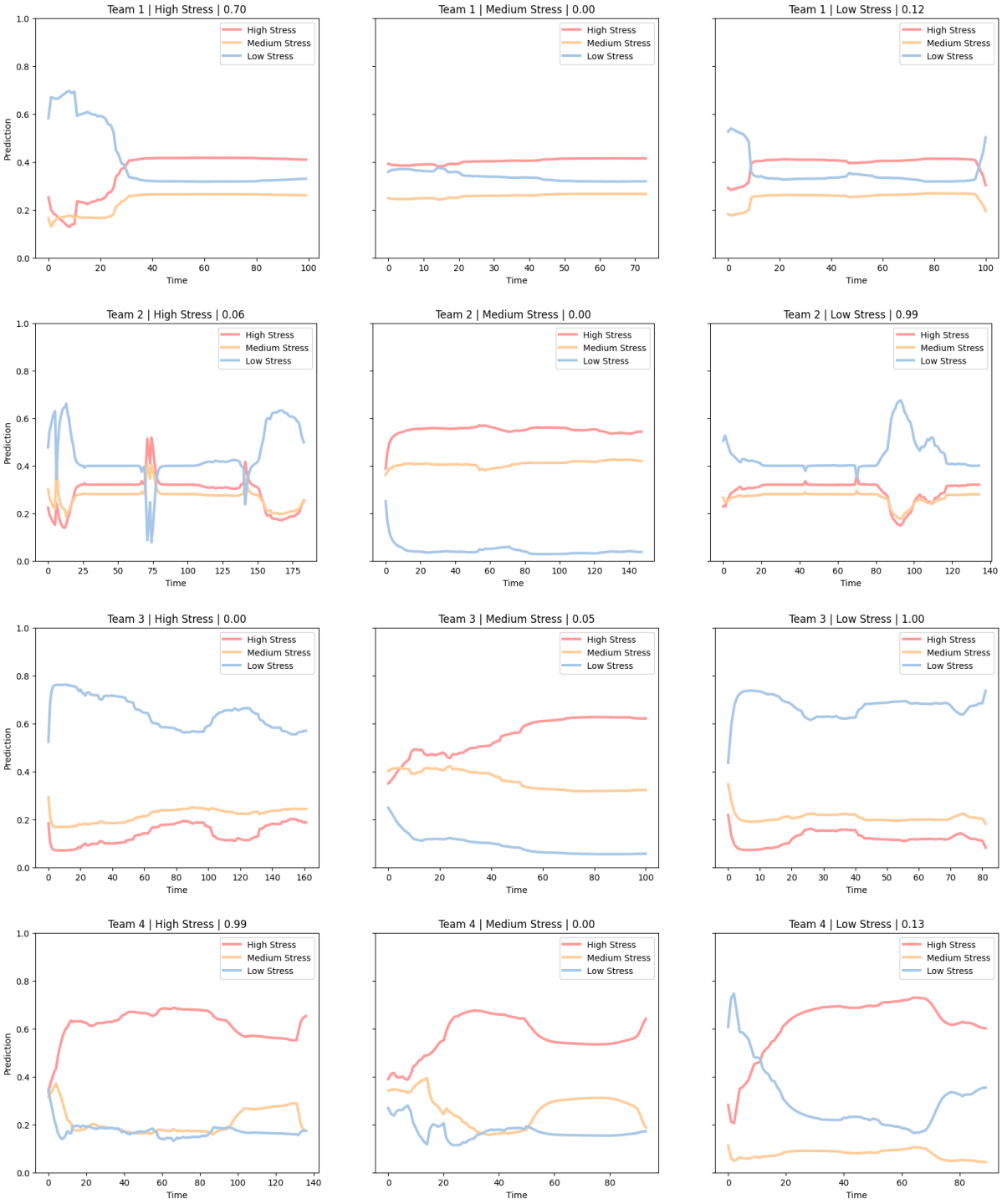
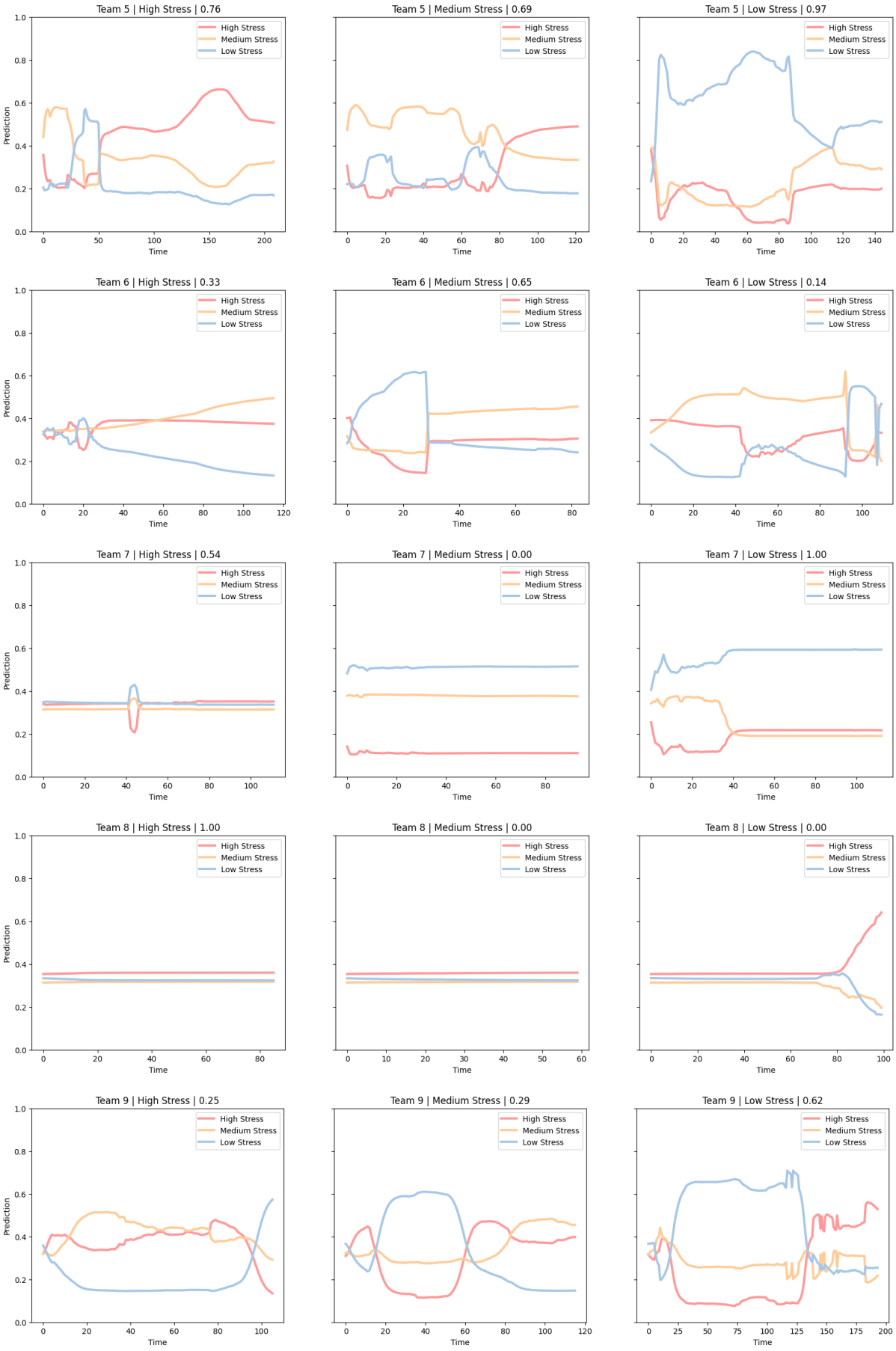
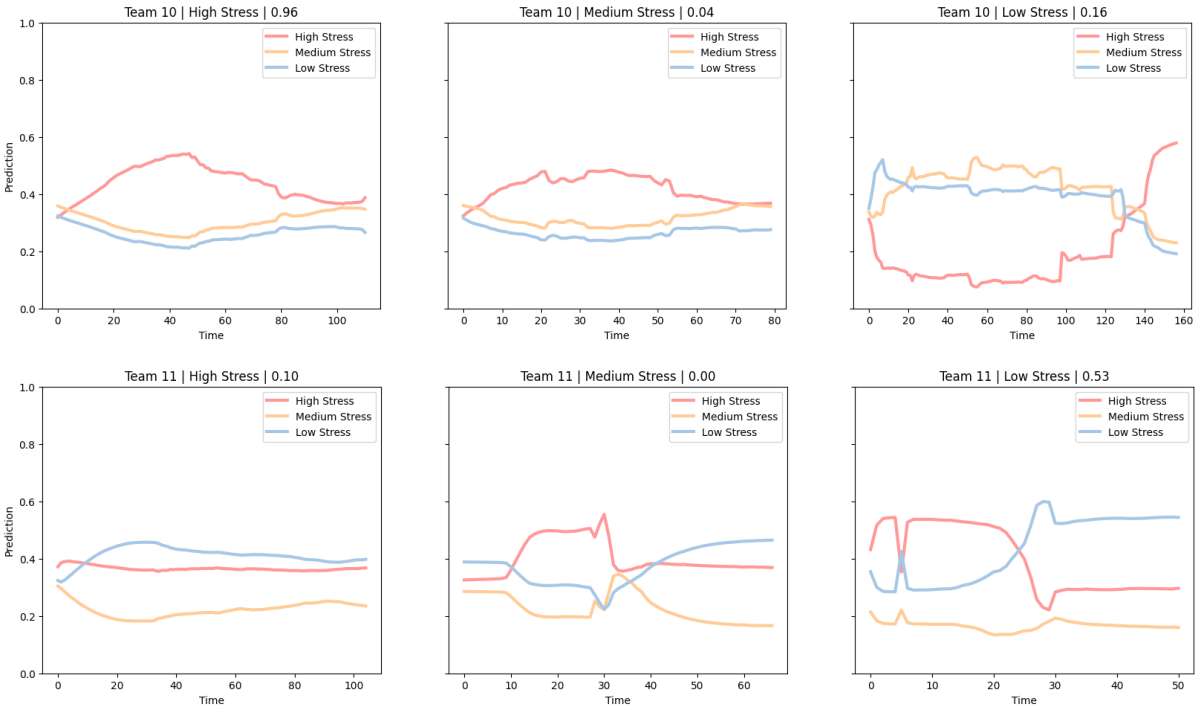

Team 8 | High Stress | 0.00
Team 8 | Medium Stress | 1.00
Team 8 | Low Stress | 1.00
Team 9 | High Stress | 0.00
Team 9 | Medium Stress | 1.00
Team 9 | Low Stress | 1.00
Team 10 | High Stress | 0.00
Team 10 | Medium Stress | 1.00
Team 10 | Low Stress | 1.00
Team 11 | High Stress | 0.00
Team 11 | Medium Stress | 1.00
Team 11 | Low Stress | 1.00
High Stress
Medium Stress
Low Stress
Prediction
Time

Figure 3. Timestep-level Scenario Prediction Probabilities using multimodal (physiology + behavior) input for each Team and Scenario




Team 5 | High Stress | 0.76
Team 5 | Medium Stress | 0.69
Team 5 | Low Stress | 0.97
Team 6 | High Stress | 0.33
Team 6 | Medium Stress | 0.65
Team 6 | Low Stress | 0.14
Team 7 | High Stress | 0.54
Team 7 | Medium Stress | 0.00
Team 7 | Low Stress | 1.00
Team 8 | High Stress | 1.00
Team 8 | Medium Stress | 0.00
Team 8 | Low Stress | 0.00
Team 9 | High Stress | 0.25
Team 9 | Medium Stress | 0.29
Team 9 | Low Stress | 0.62
High Stress
Medium Stress
Low Stress
Prediction
Time


Team 10 | High Stress | 0.96
Team 10 | Medium Stress | 0.04
Team 10 | Low Stress | 0.16
Team 11 | High Stress | 0.10
Team 11 | Medium Stress | 0.00
Team 11 | Low Stress | 0.53
High Stress
Medium Stress
Low Stress
Prediction
Time