Machine Learning for Predicting Team Functioning in HERA Missions

Shrivatsa Mishra[1], Caroline J. Wendt[1], Sydney Begerowski[2], Suzanne Bell[3], Theodora Chaspari[1]

[1]University of Colorado Boulder, CO, USA; [2] KBR, Houston, TX, USA; [3] NASA Johnson Space Center, Houston, TX, USA
shrivatsa.mishra@colorado.edu, caroline.wendt@colorado.edu, theodora.chaspari@colorado.edu,
sydney.r.begerowski@nasa.gov, suzanne.t.bell@nasa.gov

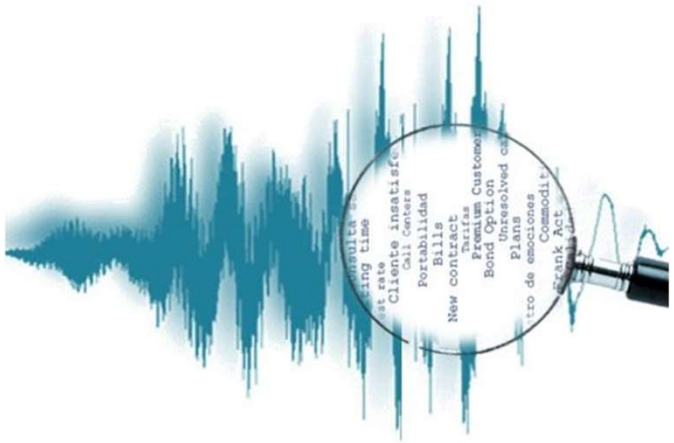
This research was funded by the NASA Human Exploration Research Opportunities (HERO) Program, #80NSSC25K7249 (PI: Chaspari, Co-I: Bell).

Motivation

NASA's Artemis and Mars missions involve extreme isolation, confinement, and communication delays.

Traditional monitoring (self-reports) is intrusive, subject to "survey fatigue," and provides "lagging" indicators of team health.

Goal: Develop an unobtrusive, real-time system using speech to detect degradations in team functioning before mission-critical failures occur.



Research aim

To develop a speech-based AI system to unobtrusively predict degradation in team functioning during NASA missions.

The target system:

- Automatically analyzes prosodic (tone of voice) and linguistic (language) content of speech
- Models their interaction among team members
- Extracts influence and turn-taking statistics among speakers, without requiring additional human coding

Study objectives

Design the AI system

Design ML models that predict team functioning using speech measures:

1. General vs team-specific models
2. Time-based vs static models
3. Ablation studies (e.g., comparison between computer-generated and manually-corrected data)

Identify the most important speech-based measures and mission components

Toward this goal, we identify:

1. A minimal subset of speech-based measures that are effective predictors of team functioning degradation for each part of the mission
2. Specific parts of the mission that primarily contribute to team functioning degradation

Data overview

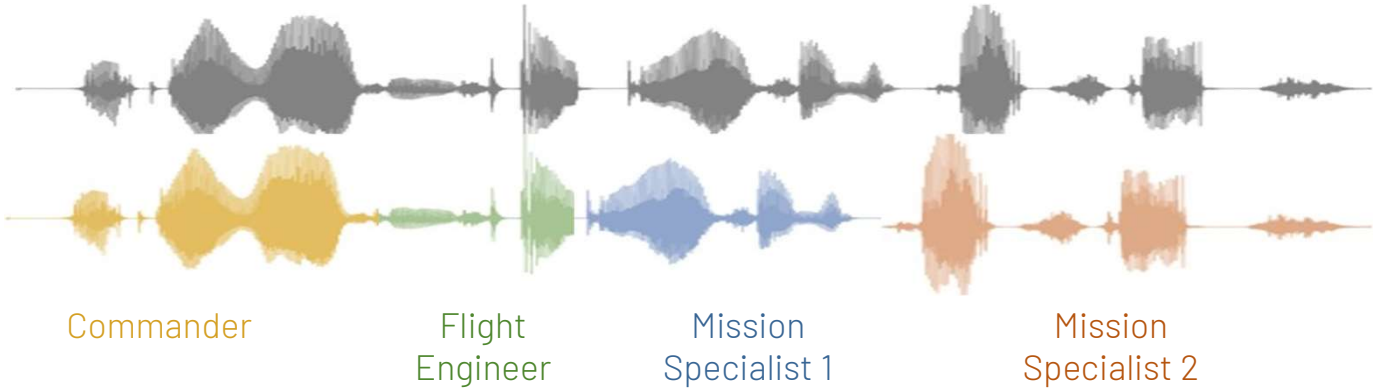
NASA HERA Campaigns 4 and 5 (C4, C5)

- 9 teams, 4 members each, 45 days
- Team interaction battery (TIB)
  - 1.5 hours, x5 per mission (x1 pre-mission, x4 in-mission)
  - Decision-making task: team must communicate distributed information to identify ideal solution to a problem
  - Relational task: team answers questions regarding personal experiences, values, and opinions
- Multi-mission space exploration vehicle – extra vehicular activity (MMSEV-EVA)
  - Operationally-relevant team performance task
  - x23 in-mission, sampled from the x5 mission days that align with TIB
  - Pre-MMSEV-EVA discussions: informal interactions before the MMSEV-EVA task of 9.91 minutes of average length

Audio pre-processing

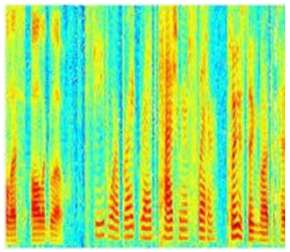
(a) Unobtrusive audio recording of astronaut team interaction

(b) Audio pre-processing: automatic transcription, voice activity detection, speaker segmentation/identification



Speech-based features

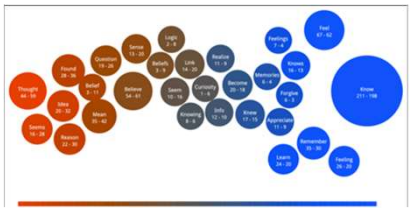
Acoustic Measures



Voice parameters related to frequency, energy/amplitude and spectral balance of speech with theoretical significance

OpenSMILE - 118 features

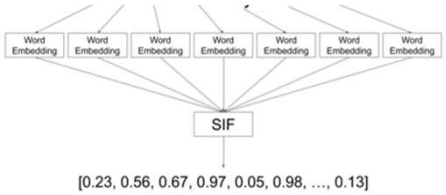
Dictionary-based language measures



Descriptors of linguistic, psychological, and cognitive processes based on a predefined dictionary

Linguistic Inquiry and Word Count (LIWC) - 88 features

Acoustic Measures



Semantic and syntactic representation of a sentence based on large-scale public language corpora

Universal Sentence Encoder - 512 features

Team functioning outcomes

All experiments were conducted as binary classification to predict low and high team functioning

Outcome 1 (objective): Task accuracy

- TIB: Correct decision (Best/Middle) (N = 25) vs Incorrect decision (N = 19)
- MMSEV-EVA: Based on %objectives met(N=30)(operationally-relevant performance)

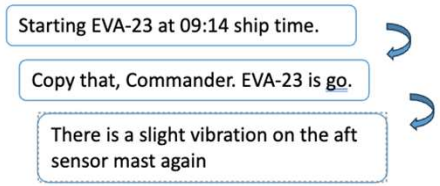
Outcome 2 (subjective): Self-reported team efficacy and cohesion

- TIB: High efficacy/cohesion (N = 17) vs Low efficacy/cohesion (N = 17) based on the median

Configurations of machine learning models

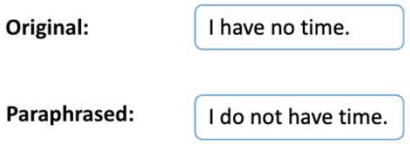
Static vs Time-based models

Static: consider each turn of the dialog independently (logistic regression, random forests)
Time-based: model the evolution of speech patterns across turns (long short-term memory neural networks – LSTM)



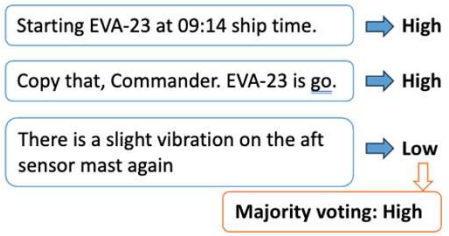
Data augmentation

Generates additional training samples by modifying existing text (e.g., synonym replacement, paraphrasing) Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002)



Aggregation across turns

Decisions obtained from the static models at the turn-level aggregated with majority voting for the entire session

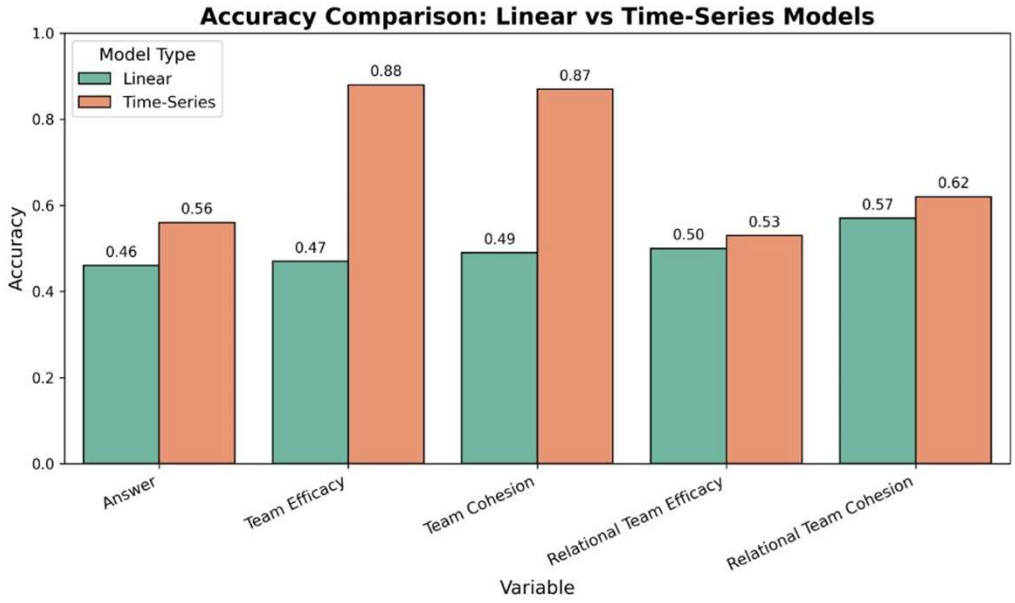


Results - Balanced Accuracy Across Team Functioning Outcomes

When running a team-independent analysis, LSTM models predicted self-reported team efficacy and team cohesion during the decision-making task significantly better as compared to the linear models reaching 87- 88%

LSTM models do not achieve significantly better performance in terms of answer accuracy, although there is an increase to 56%

LSTM models perform slightly better in the relational task compared to the static models, but the difference is not significant.

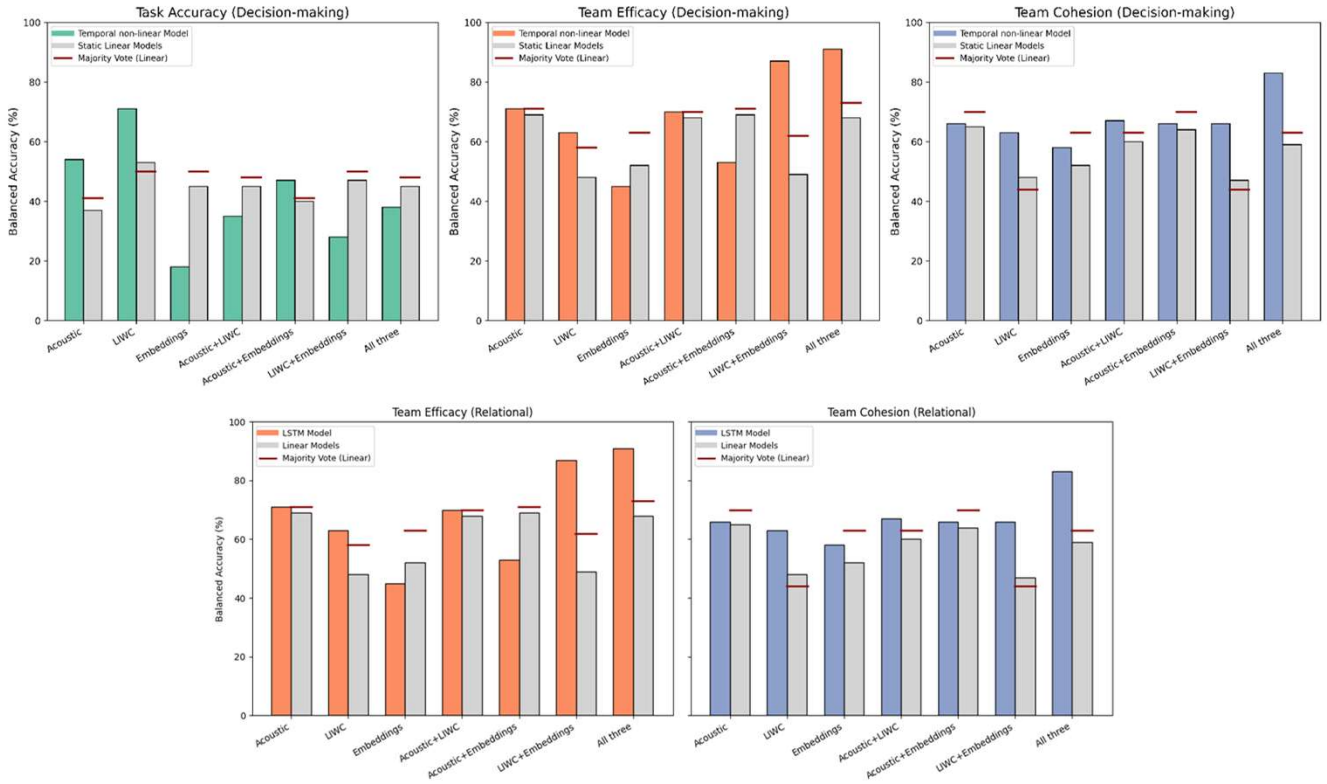


Results - TIB: Static vs time-based models (Leave-one-day-out cross-validation)

Metrics: Task Accuracy, Self-Reported Efficacy & Cohesion.

Achieved up to 90% accuracy in predicting Team Efficacy and Cohesion, while task accuracy lagged behind at 71%.

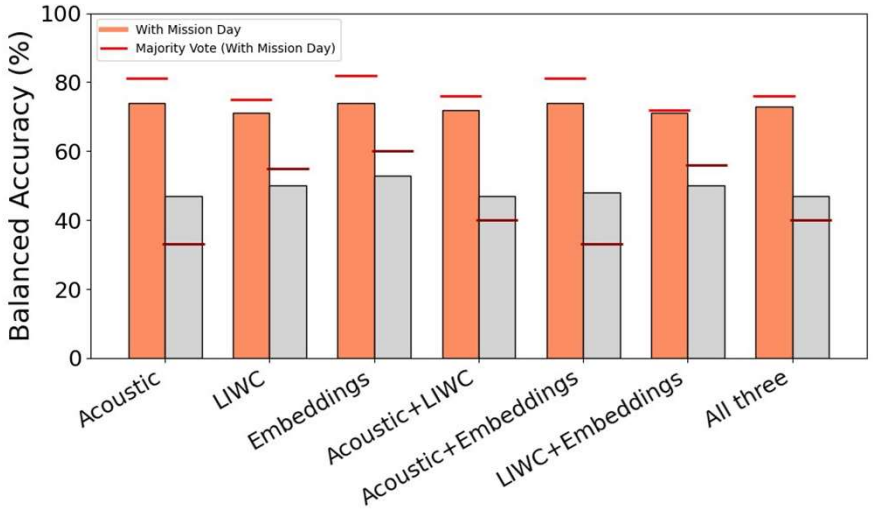
Models performed significantly better on the TIB relational task compared to the decision-making task.



Results - MMSEV: Mission day as input (Leave-one-day-out cross-validation)

Mission day added as an input variable to model the team technical knowledge retention across mission days in the MMSEV-EVA (i.e., in contrast, time-based models - LSTMs model the progression of speech patterns across turns within one day of the mission)

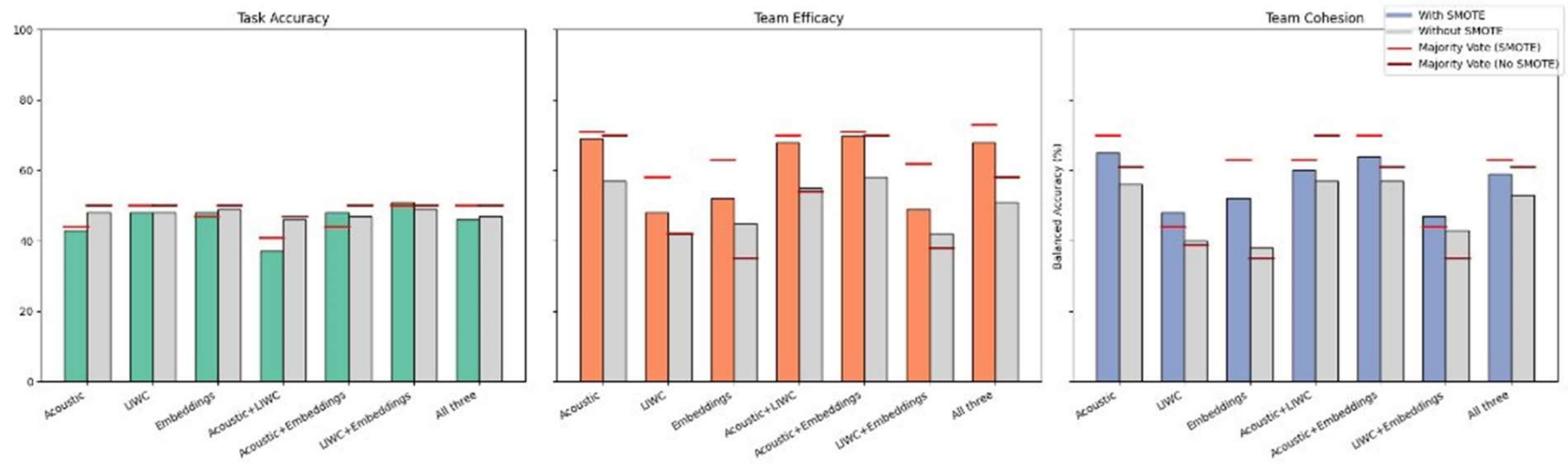
Language embeddings and acoustic features are the best performing reaching 82% accuracy, while LIWC features are slightly worse performing.



Results - TIB: Impact of data augmentation

Data augmentation generally increases performance across all data configurations for predicting self-reported team efficacy and cohesion in the decision-making part

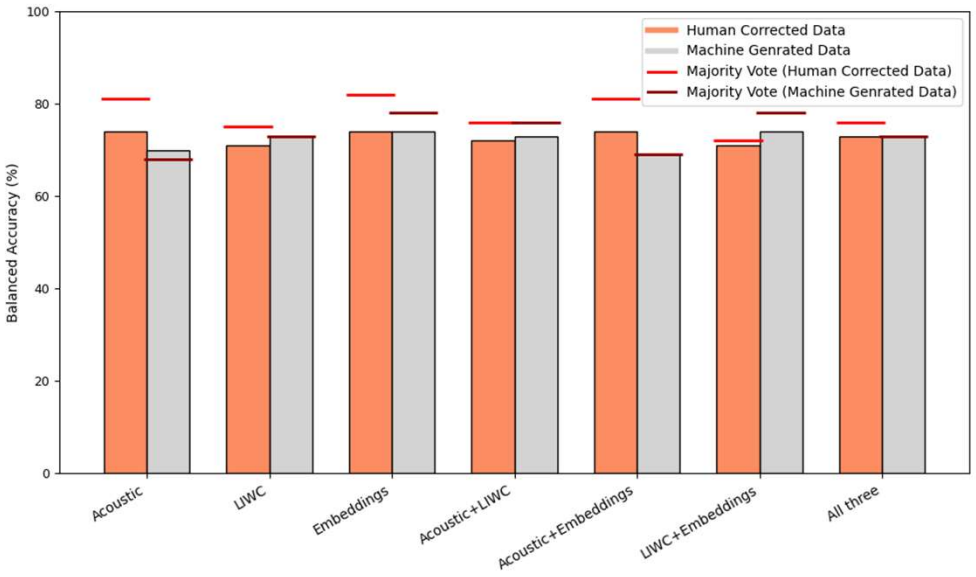
Implementing majority voting reduces the gains of data augmentation in certain cases



Results - MMSEV-EVA: Manual vs automatic speech pre-processing

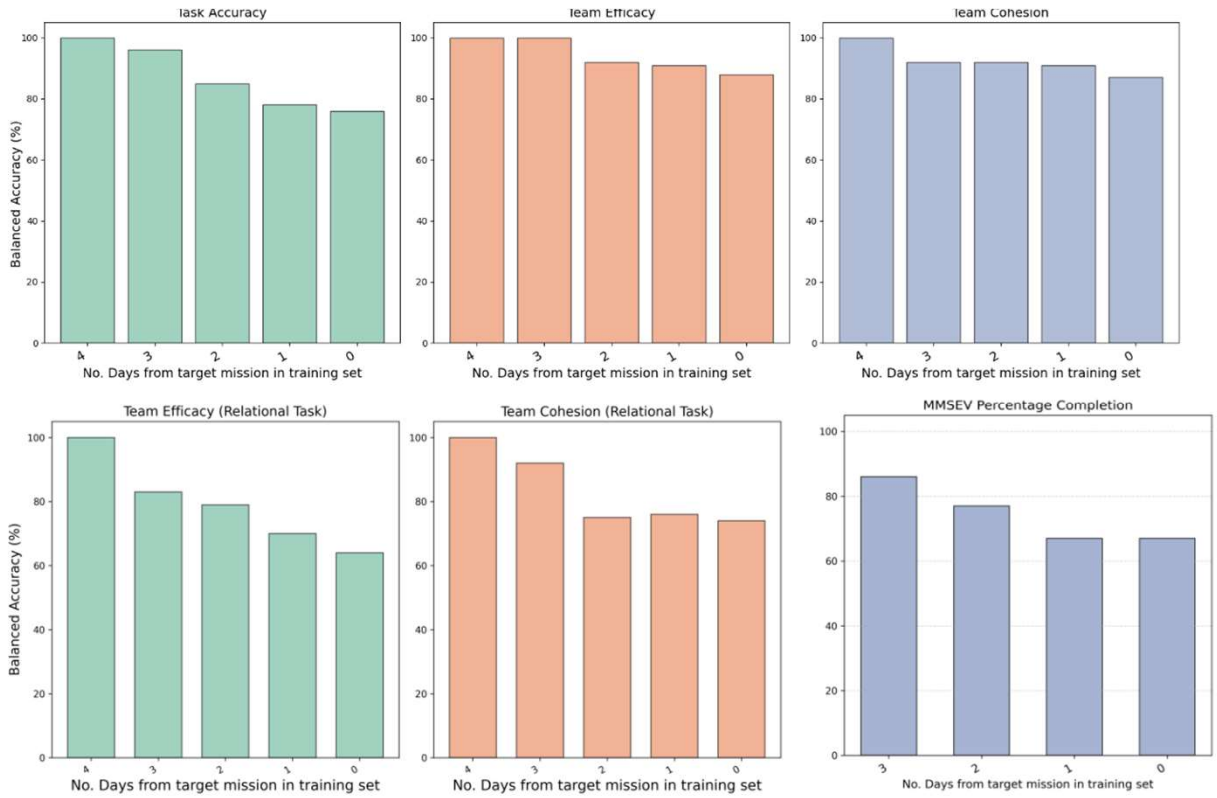
Manual vs. Machine Data: Models trained on human-corrected data showed modest performance gains, particularly when utilizing acoustic features.

Transcription Robustness: No significant correlation found between Word Error Rate (WER) and model accuracy ($r(55) = -0.08$, $p = 0.51$).



Results- TIB: Impact of team-specific data

- Data from previous days of the target team were added to the training set. Model was evaluated on the remaining days.
- Performance across all outcome measures significantly improves when team-specific information from previous days is available.



Conclusion - TIB

- Time-based models overall outperform static models
  - Task accuracy is near chance level (~50%) with static models, but increases to 56% for leave-one-team-out cross-validation (team independent) and 71% for leave-one-day-out cross-validation (team-dependent) when using time-based models
  - Self-reported efficacy and team cohesion reach 70% (decision-making task) and 78-85% (relational task) with static models, and improve to up to 82-91% with time-based models
- Data augmentation and majority voting depict model performance gains across outcomes
- Leave-one-day-out performs better than leave-one-team-out, underscoring the team-dependent nature of the interactions.

Conclusion - MMSEV

- Time-based models overall outperform static models
  - Static models reach 60% accuracy (language embeddings) on predicting task accuracy
  - Time-based models reach up to 82.9% accuracy when using acoustic features only
  - Improvement in performance of static models when adding mission day as a feature reaching 81-82% accuracy (language embeddings, acoustic features), underscoring the strong learning effect in the MMSEV-EVA
- Model performance is overall higher when trained on human-corrected data. However, models trained on machine-generated data achieve comparable performance. Significant model performance benefits are observed when the machine-generated transcripts are of high quality over those of middle and low quality

Design recommendations for NASA standards

Finding	Design recommendation
Having 1-2 days of prior data from each team improves prediction of team functioning.	Minimum requirement of at least 1-2 days of interaction data is collected prior to the target interaction.
Short data segments preceding a focal task (e.g., decision-making discussions in TIB, informal pre-task conversations in MMSEV-EVA) yield performance gains in predicting team functioning.	A brief, pre-task collection of interaction data in addition to targeted team task data collection is recommended.
Time is an important predictor of MMSEV-EVA task performance.	For long-duration missions, modelling time (and accounting for temporal practice or learning effects) is a critical for monitoring team performance on operationally relevant team performance tasks that have a practice effect.
Range of improvements was observed when designing machine learning models based on manually corrected acoustic data as compared to machine-generated ones.	Minimum-quality standards for automated tools potentially with a tiered framework (e.g., fully/semi-automated) with corresponding performance-efficiency trade-offs should be pursued.
Acoustic features and language embeddings yielded the best performance in predicting team functioning.	Complementary use of both acoustic and language representations tailored to the operational environment (e.g., focal task, team functioning outcome, recording device) is encouraged.

THANK YOU

Email: shrivatsa.mishra@colorado.edu, theodora.chaspari@colorado.edu