Alhassane Samassekou

ITAI 2372

Testing Plan: Mission Whisper - AI-Powered Astronaut Stress Monitoring System

Overview

This testing plan ensures that **Mission Whisper** accurately detects astronaut stress across voice, facial, biometric, and text data. It includes unit testing, simulation testing, and validation to meet performance and reliability requirements.

Testing Objectives

- Validate individual AI model performance for each data type.
- Ensure system-level accuracy in detecting stress states.
- Verify usability and reliability of the real-time dashboard.
- Confirm robustness under space mission conditions (e.g., noisy data).

Unit Testing

Scope

Test each AI pipeline independently:

- **Voice Emotion Recognition**: CNN/transformer model on audio inputs.
- **Facial Expression Analysis**: ResNet model on video frames.
- **Biometric Stress Classification**: Random Forest model on physiological data.
- **Text Sentiment Analysis**: BERT model on mission logs.

Test Cases

- **Voice**: Input labeled audio clips (e.g., RAVDESS dataset) with known emotions (e.g., stressed, calm). Verify model predictions match labels.
- **Facial**: Use CK+ dataset with annotated expressions (e.g., tense, neutral). Check classification accuracy.
- **Biometric**: Feed synthetic heart rate/skin temperature data simulating stress vs. rest. Confirm correct stress level classification.

- **Text**: Analyze sample logs with positive, negative, and neutral sentiments. Validate sentiment scores.

Metrics

- Accuracy, precision, recall, F1 score for each model.
- Confusion matrices to evaluate multi-class performance (e.g., low/medium/high stress).

Simulation Testing

Scope

Test the integrated system under realistic mission scenarios.

Test Cases

- **Scenario 1**: Simulate a high-stress event (e.g., equipment failure) with audio, video, biometric, and text inputs. Verify risk score and alert generation.
- **Scenario 2**: Inject noisy or incomplete data (e.g., audio dropouts, missing biometric readings). Ensure system handles errors gracefully.
- **Scenario 3**: Simulate normal operations with low stress. Confirm no false-positive alerts.

Metrics

- System-level accuracy in detecting stress states.
- False positive/negative rates for alerts.
- Response time for real-time processing.

Validation Approach

- **Human-in-the-Loop**: Psychologists review model predictions on synthetic and historical datasets to ensure clinical relevance.
- **Threshold Tuning**: Adjust alert thresholds based on feedback to balance sensitivity and specificity.
- **Cross-Validation**: Use k-fold cross-validation during model training to ensure generalizability.
- **Stress Testing**: Simulate edge cases (e.g., extreme biometric readings) to verify system stability.

Evaluation Metrics

- **Primary Metrics**: Accuracy (>85%), F1 score (>0.80) for stress detection.
- **Secondary Metrics**: Dashboard refresh rate (<1s), alert latency (<2s).
- **Qualitative**: User feedback from psychologists on dashboard usability and alert clarity.

Testing Tools

- **Python**: Unit test frameworks (e.g., pytest).
- **Datasets**: RAVDESS (audio), CK+ (facial), synthetic biometric datasets.
- **Simulation**: Custom scripts to generate mission-like data streams.

Validation Timeline

- **November 2024**: Conduct unit tests on individual models.
- **Early December 2024**: Run simulation tests and human-in-the-loop validation.
- **December 10, 2024**: Finalize testing results for submission.

Notes

- Testing will prioritize real-world applicability, focusing on noisy or incomplete data scenarios common in space.
- Results will be documented with detailed metrics and visualizations (e.g., confusion matrices).