Silverfish Safe Passage System Evaluation

Human and AI Intervention

Sanika Mhaudgat, Krutin Rathod, Carlo Lipizzi, Devam Mondal

## **Abstract:**

In evaluating the Silverfish Safe Passage System, we aimed to determine the necessity of human intervention by analyzing various factors that influence AI and human accuracy. Through a series of statistical tests, including paired t-tests and ANOVA, we assessed the performance of both AI and human operators across different terrains and environmental conditions.

Our findings indicate that AI generally maintains high accuracy in "Grassy" and "Sandy" terrains, with some variability in "Swampy" and "Wooded" terrains. However, significant performance drops were observed in "Rocky" and "Wooded" terrains for both AI and humans. Environmental conditions such as visibility and precipitation also impacted detection accuracy. Based on these insights, we recommend developing specialized detection techniques for challenging terrains and enhancing AI training with diverse data to improve robustness.

#### **Map A Day - Analysis:**

Our initial examination of performance between humans and AI involves the data from Map A Day. This dataset reflects accuracy for both AI and human participants on a 10x10 grid featuring different terrains and includes two distinct locations tested at various times. Here are some notes that interpret our analysis of the dataset.

**AI Performance Analysis:**

* AI accuracy is generally high, with most values around 0.95 to 0.96.
* There are drops in performance (e.g., 0.56-0.57) which may correspond to certain terrain types or environmental conditions.
* Notable low-performance areas:
  + Rows 4-7, Column 2: Accuracy drops to 0.56, possibly due to "Wooded" terrain.
  + Rows 3-6, Column 4: Accuracy ranges between 0.62 and 0.70, indicating potential challenges with "Rocky" terrain.
* High accuracy is maintained in most of the "Grassy" and "Sandy" terrains, with some variability in "Swampy" and "Wooded" terrains.

**Human Performance Analysis:**

* Human accuracy is slightly lower than AI, mostly ranging from 0.85 to 0.90.
* There are significant drops in performance in specific conditions:
  + Row 4, Column 2: Accuracy drops to 0.56, aligning with AI performance issues in "Wooded" terrain.
  + Rows 5-7, Columns 1-3: Accuracy drops to 0.57-0.75, indicating challenges with "Wooded" and "Rocky" terrains.
* Generally, humans perform better or similarly in "Grassy" and "Sandy" terrains compared to more challenging "Wooded" and "Swampy" terrains.

**Terrain and Performance Correlation:**

* **Grassy Terrain**: Both AI and humans maintain high accuracy, with minimal performance drops.
* **Rocky Terrain**: Significant performance drops for both AI and humans, indicating this terrain poses challenges.
* **Sandy Terrain**: High accuracy for both AI and humans, similar to "Grassy" terrain.
* **Wooded Terrain**: Notable drops in performance for both AI and humans, suggesting dense vegetation impacts detection abilities.
* **Swampy Terrain**: Mixed results, with performance generally lower than "Grassy" and "Sandy" but better than "Rocky" and "Wooded."

**Environmental Conditions:**

* The test times (1000, 2200 for Location A and 0900, 2100 for Location B) could affect visibility and temperature, influencing performance.
* Current conditions (e.g., 70°F, 5 mph wind, 0.05 visibility, 5 precipitation) suggest moderate challenges, with visibility and precipitation potentially impacting detection accuracy.

**Recommendations:**

* **Terrain-Specific Strategies**: Implement specialized detection strategies tailored for "Rocky" and "Wooded" terrains to enhance accuracy.
* **Environmental Adjustments**: Integrate factors such as time of day and weather conditions into the planning of detection operations to optimize performance.
* **Model Training**: Strengthen AI training by incorporating a broader range of terrain and environmental data, increasing its robustness in challenging scenarios.

This analysis establishes a framework for understanding the interplay between terrain types, environmental conditions, and system performance in mine detection tasks. Continued research and model refinement are essential for further improving accuracy and reliability.

#### **Map A Night - Analysis:**

Based on the second set of provided data, Map A Night, we assessed AI and human performance in mine detection across different terrains:

**Performance Data Summary:**

* **AI Performance**: Varies between 44% and 75% accuracy.
* **Human Performance**: Varies between 70% and 85% accuracy.

**Paired t-Test Results:**

We also performed a paired t-Test to evaluate the difference between AI and human performance. The following are our findings. For detailed information, please refer to the “Statistical Tests” section.

* **Grassy Terrain**: AI performance is notably inferior to human performance.
* **Rocky Terrain**: Human performance significantly surpasses AI performance.
* **Sandy Terrain**: AI performance significantly exceeds human performance.
* **Swampy Terrain**: AI demonstrates a substantial advantage over human performance.
* **Wooded Terrain**: Human performance is considerably better than AI performance.

**Two-Way ANOVA Results**:

We performed a two-way ANOVA test to evaluate the difference between AI and human performance. The following are our conclusions. For further details, please refer to the “Statistical Tests” section.

* **Surface Type**: There is a highly significant impact on performance, with substantial variation across different terrain types.
* **Detection Method**: The test indicates a significant effect on performance, highlighting a difference between AI and human detection abilities.
* **Interaction Effect (Surface Type \* Method)**: This effect is highly significant, demonstrating that the disparity between AI and human performance is influenced by the terrain type.

### **Performance Analysis by Terrain and Conditions:**

We also evaluated how performance varies across different terrains and conditions. Our conclusions are as follows:

1. **Grassy Terrain**:

* + **AI Accuracy**: Typically around 75%.
  + **Human Accuracy**: Typically around 85%.
  + **Inference**: Humans significantly outperform AI in grassy terrain.

2. **Rocky Terrain**:

* + **AI Accuracy**: Ranges from 44% to 75%.
  + **Human Accuracy**: Generally around 70%.
  + **Inference**: Humans significantly outperform AI in rocky terrain.

3. **Sandy Terrain**:

* + **AI Accuracy**: Ranges from 44% to 75%.
  + **Human Accuracy**: Generally around 70%.
  + **Inference**: AI significantly outperforms humans in sandy terrain.

4. **Swampy Terrain**:

* + **AI Accuracy**: Exhibits exceptional performance, often exceeding that of humans.
  + **Human Accuracy**: Generally around 70%.
  + **Inference**: AI holds a significant performance advantage in swampy terrain.

5. **Wooded Terrain**:

* + **AI Accuracy**: Generally ranges from 44% to 75%.
  + **Human Accuracy**: Typically around 70%.
  + **Inference**: Humans significantly outperform AI in wooded terrain.

### **Conclusion and Recommendations:**

* **AI Strengths**: AI demonstrates outstanding performance in swampy and sandy terrains. Optimizing AI algorithms for these environments could significantly enhance detection efficiency and effectiveness.
* **Human Strengths**: Humans excel in grassy, rocky, and wooded terrains. Utilizing human expertise in these areas can lead to improved detection accuracy.
* **Terrain-Specific Strategies**:
  + **Grassy and Sandy Terrains**: Prioritize AI deployment to leverage its higher accuracy.
  + **Rocky and Wooded Terrains**: Rely on human experts to take advantage of their superior performance.
  + **Swampy Terrain**: Employ AI for its exceptional detection capabilities in these conditions.

#### **Map B Day - Analysis:**

Based on the second dataset, "Map B Day," we evaluated AI and human performance in mine detection across various terrains. Below are the key findings:

**Performance Data Summary:**

* **AI Performance**: Accuracy ranges from 53% to 90%.
* **Human Performance**: Accuracy ranges from 75% to 90%.

**AI Performance Table Highlights**:

* **High Accuracy**: AI achieves high accuracy in some terrains, reaching up to 90% in several cases.
* **Low Accuracy**: AI shows significantly lower accuracy (down to 53%) in certain conditions, especially in rocky and wooded terrains.

**Human Performance Table Highlights**:

* **High Accuracy**: Human performance generally maintains high accuracy, up to 90%, across most terrains.
* **Consistency**: Human performance is more consistent than AI, with fewer significant drops in accuracy.

**Surface Type Table Highlights**:

* **Variety of Terrains**: The table displays a range of terrains, including Wooded, Grassy, Sandy, Rocky, and Swampy, spread across different rows and columns.

**Conditions**:

* **DateTime**: 0900 (Morning).
* **Temperature**: 75°F (Moderate).
* **Wind Speed**: 2 mph (Low).
* **Visibility**: 0.15 miles (Moderate).
* **Precipitation**: 5 inches (Moderate).

### **Inferences:**

1. **Terrain Influence**:

* + **Wooded Terrain**: AI accuracy is lower in wooded areas compared to human performance, which remains consistently higher.
  + **Grassy Terrain**: Both AI and human performances are relatively strong, though human performance is more stable.
  + **Sandy Terrain**: AI performs well with high accuracy, often matching or exceeding human performance.
  + **Rocky Terrain**: AI experiences difficulties with lower accuracy, while human performance stays relatively high.
  + **Swampy Terrain**: AI demonstrates very high accuracy, outperforming human performance in this terrain.

2. **Conditions Impact**:

* + **Morning (0900)**: Morning conditions, with better visibility and lower wind speeds, may enhance both AI and human performance, contributing to generally higher accuracy rates compared to nighttime conditions.

### **Recommendations:**

1. **Terrain-Specific Approaches**:

* + **Wooded and Rocky Terrains**: Consider enhancing AI algorithms or relying on human expertise to boost detection accuracy.
  + **Sandy and Swampy Terrains**: Deploy AI due to its superior performance in these environments.

2. **Condition Adaptation**:

* + **Adjust for Time of Day**: Ensure that both AI and human operators are aware of how conditions affect performance and adjust strategies accordingly. For example, AI may perform differently under varying visibility or weather conditions.

3. **Consistency and Reliability**:

* + **AI**: Prioritize training and optimizing algorithms for terrains where AI currently underperforms.
  + **Humans**: Leverage human expertise in terrains where they consistently achieve better results than AI.

By addressing these factors, teams can develop targeted strategies to enhance mine detection performance, effectively utilizing the strengths of both AI and human experts across various terrains and conditions.

#### 

#### **Map B Night - Analysis:**

Using the third dataset, "Map B Night," we evaluated AI and human performance in mine detection across various terrains. Below are our key findings:

**AI Performance Table (% Accuracy)**:

* **Accuracy Range**: 48% to 83%
* **Notable Trends**: While AI demonstrates high accuracy in certain conditions, there are significant drops in some rows and columns.

**Human Performance Table (% Accuracy)**:

* **Accuracy Range**: 70% to 80%
* **Notable Trends**: Human performance is consistently high with less variability compared to AI.

**Surface Type Table**:

* **Variety**: The table includes Wooded, Grassy, Sandy, Rocky, and Swampy terrains spread across the grid.

**Conditions**:

* **DateTime**: 2100 (Nighttime)
* **Temperature**: 60°F (Moderate)
* **Wind Speed**: 3 mph (Low)
* **Visibility**: 0.15 miles (Moderate)
* **Precipitation**: 5 inches (Moderate)

### **Analysis and Inferences:**

1. **AI Performance Trends**:

* + **High Accuracy**: AI achieves high accuracy (up to 83%) in certain conditions, especially in grids with Sandy or Swampy terrains.
  + **Low Accuracy**: AI performance significantly drops (down to 48%) in specific cells, indicating potential challenges with certain terrains or environmental conditions.

2. **Human Performance Trends**:

* + **Consistent Accuracy**: Human performance remains relatively stable and high (70% to 80%) across different conditions, suggesting reliability regardless of terrain or time of day.

3. **Surface Type Impact**:

* + **Wooded and Rocky Terrains**: AI shows lower accuracy in these terrains, while human performance remains relatively consistent.
  + **Grassy, Sandy, and Swampy Terrains**: AI performs better, particularly in Sandy and Swampy conditions.

4. **Impact of Conditions**:

* + **Nighttime (2100)**: Reduced visibility compared to daytime may contribute to the variability in AI performance, particularly in mine detection.
  + **Precipitation**: Moderate precipitation may affect both AI and human performance, though its impact appears less significant compared to visibility and time of day.

### **Recommendations:**

1. **Terrain-Specific Optimization**:

* + **Wooded and Rocky Terrains**: Improve AI algorithms or integrate AI with human oversight to enhance accuracy.
  + **Sandy and Swampy Terrains**: Capitalize on AI’s strengths, as it performs better in these environments.

2. **Adapting to Conditions**:

* + **Visibility and Time of Day**: Factor in the time of day and visibility when deploying AI systems. Enhanced algorithms or additional sensors may be necessary for effective nighttime operations.

3. **Ensuring Consistency**:

* + **Human Operators**: Continue to rely on human operators for consistent performance, especially in terrains where AI is less effective.

By focusing on these areas, teams can optimize mine detection strategies, improve the synergy between AI and human expertise, and effectively adapt to different environmental conditions.

#### **Statistical Tests:**

To support our observations and key findings, we performed statistical tests to determine if there were significant differences in AI and human accuracy. The detailed results of these tests are provided below.

#### **Map A Day - Statistical Tests:**

To statistically compare the performance of AI and human experts in mine detection across various terrain types, the following steps were taken:

1. **Formulate Hypotheses:**
   * Null Hypothesis (): There is no significant difference in performance between AI and human experts.
   * Alternative Hypothesis (): There is a significant difference in performance between AI and human experts.
2. **Choose the Test:**
   * Given that we are comparing two independent samples (AI and human accuracy) across multiple conditions, a suitable test is the paired t-test for each terrain type.
   * Alternatively, a two-way ANOVA can be used to assess the interaction between terrain types and detection methods (AI vs. humans).
3. **Prepare the Data:**
   * The performance data for AI and human experts were organized by terrain type.
4. **Results:**

{'Grassy': (18.379275936051393, 1.6005584017581558e-17),

'Rocky': (-9.359191741068397, 3.971772033676578e-09),

'Sandy': (30.833655081976268, 4.947724001565193e-12),

'Swampy': (inf, 0.0),

'Wooded': (-12.326902563068609, 3.2700143621959085e-10)}

#### **Inference from Paired t-Test Results:**

The paired t-test results compare the performance of AI and human experts across various terrain types. The test statistics (t-values) and p-values for each terrain type are as follows:

1. **Grassy Terrain**:
   * t-value: 18.379
   * p-value: 1.60e-17
   * **Inference**: The exceptionally high t-value and very low p-value suggest a significant difference in performance between AI and human experts in grassy terrain. The positive t-value indicates that AI performs significantly better than humans in this terrain type.
2. **Rocky Terrain**:
   * t-value: -9.359
   * p-value: 3.97e-09
   * **Inference**: The large negative t-value and extremely low p-value suggest a significant difference in performance between AI and human experts in rocky terrain. The negative t-value indicates that human performance is significantly better than AI performance in this terrain type.
3. **Sandy Terrain**:
   * t-value: 30.834
   * p-value: 4.95e-12
   * **Inference**: The exceptionally high t-value and very low p-value reveal a significant difference in performance between AI and human experts in sandy terrain. The positive t-value indicates that AI performs significantly better than humans in this terrain type.
4. **Swampy Terrain**:
   * t-value: inf
   * p-value: 0.0
   * **Inference**: The infinite t-value and zero p-value indicate an extremely significant difference in performance between AI and human experts in swampy terrain. This suggests a fundamental disparity in detection capabilities, with AI potentially achieving perfect accuracy or human performance being exceptionally poor in this terrain.
5. **Wooded Terrain**:
   * t-value: -12.327
   * p-value: 3.27e-10
   * **Inference**: The large negative t-value and extremely low p-value highlight a significant difference in performance between AI and human experts in wooded terrain. The negative t-value indicates that human performance is significantly better than AI performance in this terrain type.

**Summary:**

* AI significantly outperforms humans in grassy and sandy terrains.
* Humans have a clear advantage over AI in rocky and wooded terrains.
* Swampy terrain reveals an extreme performance gap, indicating AI holds a substantial advantage over human performance in this environment.

#### **Inference from Two-Way ANOVA Results:**

The two-way ANOVA examines the effects of surface type, detection method (AI vs. human), and their interaction on mine detection performance. Here are the results and their interpretations:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **Sum of Squares (sum\_sq)** | **df** | **F-Statistic (F)** | **p-value (PR(>F))** |
| **C(Surface\_Type)** | 2.733772 | 4 | 862.719 | 1.337449e-120 |
| **C(Method)** | 0.027612 | 1 | 34.856 | 1.606860e-08 |
| **C(Surface\_Type)**  **(Method)** | 0.411378 | 4 | 129.822 | 3.176686e-53 |
| **Residual** | 0.150517 | 190 | NaN | NaN |

1. **Main Effect of Surface Type**:
   * **F(4, 190) = 862.719**, p-value = 1.337449e-120
   * **Inference**: The exceptionally high F-statistic and very low p-value demonstrate a highly significant main effect of surface type on detection performance. This indicates that the terrain type significantly impacts mine detection accuracy, independent of whether AI or human operators are involved.
2. **Main Effect of Method**:
   * **F(1, 190) = 34.856**, p-value = 1.606860e-08
   * **Inference**: The high F-statistic and very low p-value suggest a significant main effect of the detection method on performance. This indicates a significant difference in mine detection accuracy between AI and human experts, independent of the terrain type.
3. **Interaction Effect (Surface Type \* Method)**:
   * **F(4, 190) = 129.822**, p-value = 3.176686e-53
   * **Inference**: The high F-statistic and very low p-value indicate a significant interaction effect between surface type and detection method. This suggests that the difference in detection performance between AI and humans varies significantly depending on the terrain type.

**Summary:**

* **Surface Type**: There is a highly significant impact of surface type on detection performance, indicating that the terrain type greatly influences the effectiveness of mine detection.
* **Detection Method**: There is a notable difference in detection performance between AI and human experts, suggesting that one method generally outperforms the other.
* **Interaction Effect**: The performance difference between AI and human experts varies significantly across different terrains, indicating that the effectiveness of each method depends on the specific terrain type.

These findings highlight the importance of terrain-specific strategies in mine detection. For instance:

* **Grassy and Sandy Terrains**: AI may be the preferred option due to its superior performance.
* **Rocky and Wooded Terrains**: Human expertise might be more reliable.
* **Swampy Terrains**: AI demonstrates a significant advantage.

Developing and implementing specialized detection techniques tailored to each terrain type can improve overall mine detection effectiveness.

#### **Map A Night - Statistical Tests:**

As shown above, we followed the same procedures, and the statistical tests for Map A Night are attached.

#### **Paired t-Test Results:**

* **t-Statistic:** -24.043
* **p-Value:** 4.03e-43

#### **Interpretation:**

1. **t-Statistic**:
   * The negative t-statistic suggests that, on average, AI performance is lower than that of human experts across the various surface types.
2. **p-Value**:
   * The extremely low p-value (4.03e-43) is far below the common alpha level of 0.05, indicating that the difference in performance between AI and human experts is statistically significant.

**Conclusion:**

* **Significant Difference**: The paired t-test results reveal a highly significant difference between AI and human performance, with the negative t-statistic indicating that human experts outperform AI across the various surface types examined.
* **Implication**: Since human experts consistently surpass AI in mine detection, it may be advantageous to focus on strategies that integrate human expertise with AI or to explore further enhancements and training for AI systems to improve their accuracy.

This notable difference underscores the importance of carefully evaluating AI methods and considering potential improvements to close the performance gap between AI and human experts.

#### **Map B Day - Statistical Tests:**

As demonstrated above, we followed the same procedures, and the statistical tests for Map B Day are attached.

#### **Paired t-Test Results:**

* **t-Statistic:** -1.755
* **p-Value:** 0.113

#### **Inference:**

The paired t-test yielded a t-statistic of -1.755 and a p-value of 0.113. Since this p-value exceeds the common significance level of 0.05, we do not have sufficient evidence to reject the null hypothesis.

**Interpretation:**

* **Null Hypothesis (H₀):** There is no significant difference in accuracy between AI and human performance.
* **Alternative Hypothesis (H₁):** There is a significant difference in accuracy between AI and human performance.

Given the p-value is greater than 0.05, we cannot reject the null hypothesis, suggesting that there is no statistically significant difference in performance between AI and human methods based on the current data. However, this does not mean the performances are identical; rather, any observed differences may be due to random variability rather than a systematic effect.

#### **Map B Night - Statistical Tests:**

As demonstrated above, we followed the same procedures, and the statistical tests for Map B Night are attached.

#### **Paired t-Test Results:**

* **t-Statistic:** -2.149
* **p-Value:** 0.034

**Interpretation:**

* **Statistical Significance:** The p-value of 0.034 is below the conventional alpha level of 0.05, indicating a statistically significant difference between AI and human performance.
* **Direction of Difference:** The negative t-statistic indicates that, on average, AI performance is lower than that of humans.

**Conclusion:**

The test results reveal a statistically significant difference in performance between AI and human methods, with humans outperforming AI based on the observed data. This suggests that, under the given conditions, human methods are more effective than AI in terms of accuracy. This finding may warrant further investigation into the reasons behind AI's underperformance and could inform efforts to enhance its performance.