_Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Simulation and Classification of Spatial Disorientation in a Flight use-case using Vestibular Stimulation

JAMILAH FOUCHER[1], Member, IEEE, ANNE-CLAIRE COLLET[5], KEVIN LE GOFF[3], THOMAS RAKOTOMAMONJY[2], VALERIE JUPPET[4], THOMAS DESCATOIRE[4], JERÉMIE LANDRIEU[1], MARIELLE PLAT-ROBAIN[3], FRANÇOIS DENQUIN[1,2], ARTHUR GRUNWALD[6], JEAN-CHRISTOPHE SARRAZIN[2], and BENOÎT G. BARDY[1]

[1]EuroMov Digital Health in Motion, Univ Montpellier and IMT Mines Ales, Montpellier 34090 France (benoit.bardy@umontpellier.fr)
[2]DTIS, ONERA, Salon de Provence 13300 France
[3]AIRBUS, Toulouse 31000 France
[4]AIRBUS Helicopters, Marignane 13700 France
[5]Human Design Group, Toulouse 31000 France
[6]Technion, Israel Institute of Technology 32000 Israel

Corresponding authors: Jamilah Foucher (e-mail: j622amilah@gmail.com)

This project was supported by the French Department of Civil Aviation (DGAC) and by the European Union (FEDER iMOSE).

ABSTRACT In aeronautics, spatial disorientation (SD) is "an erroneous sense of one's position and motion relative to the plane of the earth's surface" [1]. SD has a wide range of situations and factors, but mainly it has been studied using reduced experimental contexts such as motion detection experimentation in isolation. Because there are many SD use-cases that are studied in isolation in a reduced manner, it is difficult to develop a generalized and fundamental understanding of the occurrence of SD and viable solutions. To investigate SD in a generalized manner, a two-part neuroergonomics study consisting of an in-flight piloting use-case experiment and machine learning (ML) model prediction was performed. The first part of the study was the creation of a generalized SD perception dataset using whole-body experimental motion detection methods in a naturalistic flight context; participant perceptual joystick response was measured during rotational or translational vestibular stimulation. The second part of the study consisted of ML parameter tuning selection for SD prediction, using joystick response-derived features from the generalized SD perception dataset. Additional measures of SD were investigated for future ML feature usage, such as questionnaire-based physical disorientation measured using the simulator sickness questionnaire (SSQ) disorientation sub-scale. The perceptual SD dataset was statistically proven to be representative of human motion detection behavior, demonstrating that the simulation environment was sufficient to generate a fidel SD context. ML modeling comparison analysis demonstrated that SD can be accurately predicted regardless of the feature quantity used, however model type, specialized dataset models, feature type, and label type significantly influence prediction accuracy. Finally, no significant relationship between physical disorientation and motion detection was found, indicating that two-sample before and after SSQ questionnaire-based methods are insufficient to uncover correlations with perceptual disorientation; a more frequent physical disorientation measure is needed.

INDEX TERMS Aircraft navigation, human computer interaction, joystick perceptual response, machine learning algorithms, motion detection, spatial disorientation, vestibular dead reckoning.

I. INTRODUCTION

Spatial disorientation (SD), in aviation, is the failure to perceive orientation, position, or movement. It is caused by multiple factors including environmental references and conditions, experience, and stress. There are diverse types of SD symptoms, ranging from confusion to physical sickness, and currently there is no proven method or solution to prevent it [2], [3], [4], [1], [5], [6]. International

studies on the frequency and severity of SD accidents show that the cause of 6-32% of major accidents are due to SD, similarly 15-26% of fatal accidents are a result of SD [6]. Recovery from SD is strongly connected to the pilot's awareness of the situation, and his/her ability to perform corrective control, despite the disorientation, to maintain aerodynamic stability; 80% and 20% of SD incidents are caused by unrecognized and recognized situations respectively [2]. Currently, SD is treated by educating pilots about the signs and symptoms of SD and instructing them to fly below physiological thresholds of the human vestibular system [5]. Treating SD has been challenging because it is often defined with respect to a specific aeronautical context [1], [6]. SD definitions focus on flight performance errors but seldom include context-independent behavior, perceptual, or physiological trends. Because SD has been studied case by case in an aeronautical context, there is little general understanding of the onset of SD and orientation, position, or movement perception with respect to environmental references. It would be of interest to study SD using a motion detection experimental paradigm, measuring SD in a general context with respect to whole-body orientation, position, speed, and perceptual feedback. A goal to understand overall SD would be to outline a general framework for modeling and predicting the occurrence of SD based on perceptual feedback..

Vibration or motion, measured by the human vestibular system, contains important information about the environment and our orientation and position with respect to the environment. Motion detection is the act of discerning self-motion with respect to a reference in the environment [7]. Human motion detection and perception are quantified by stimulating the vestibular system systematically via different vibrational and motion experimental paradigms [8]. Initially, motion detection was quantified by observing at which directions and speeds/accelerations, angular or linear, humans could perceive self-motion. The observed values where humans could not perceive correct self-motion were called vestibular thresholds or motion detection thresholds. Movement speed and acceleration influence motion perception. Earlier motion detection studies targeted aviation applications, where thresholds were often reported in terms of acceleration because flight instrumentation and behavioral interpretation was more accessible in terms of acceleration than speed [9]. Whereas recent motion detection studies use robotic motion simulation and often report thresholds in terms of speed because robotic motion planning is more reliable in terms of speed than acceleration [10], [11], [12], [13]. Both speed and acceleration motion detection thresholds are comparable because they are directly related with the derivative or integral function. Earlier experimental paradigms included the usage of different experimental conditions such as magnitude and frequency of speed or acceleration motion stimuli trajectory, sequence and exposure time of movement and non-movement events, movement direction with respect to the orientation of the head, and whole-body stimulation [9]. Recent motion perception research has adopted robotic simulation tools and standardized experimental paradigms, including a greater range of test motion frequencies, allowing for more precise and consistent motion detection boundaries for a large variety of perceptual situations. Additionally, vestibular motion perception studies investigate context-driven parameters, such as 1) position and acceleration stimuli trajectory, direction, and rate; 2) vestibular dysfunction vs control detection; 3) orientation and/or movement of the user's body during exposure to stimuli; 4) expertise vs novice detection; and 5) age. Depending on the context parameters and the stimuli trajectory, the vestibular-proprioceptive system detects motion differently, and thus, behavioral responses are different [14], [13], [11], [10], [12]. For SD applications, motion detection thresholds were used as an indicator of SD awareness [1], [5]. However, it remains uncertain how to reliably use thresholds to assist with SD in a functional aviation context.

Recent SD research has shown that it is more accurate to predict states of disorientation from physiological, performance, or movement signatures than to use vestibular thresholds. Thus, instead of applying perceptual threshold values from motion detection research to SD research, as was done in the past in aviation, SD researchers conduct motion detection experiments using realistic flight scenarios to determine which physiological, performance, and/or environmental measures provide accurate information about a pilot's SD-state. For instance, directional perception was investigated in a realistic helicopter task in which participants were asked to point towards the sky to demonstrate a non-SD state [15]. Similarly, continuous heading detection was investigated using a compensatory task in which perceived heading was measured with respect to a remembered target [16]. Most recently, the individual and interactive influences of optical and gravito-inertial stimuli during simulated low-altitude flight demonstrated the importance of sensory integration effects on height perception using joystick response [17]. These applied studies are useful and provide insightful information regarding motion perception in realistic contexts. However many of the mentioned studies, research SD per use-case instead of trying to identify SD regardless of the SD use-case context or in a generalized manner.

In this study, we investigated SD using: 1) motion perception experimental methods to create a generalized SD occurrence dataset containing a perceptual feedback measure, and 2) statistical and machine learning (ML) methods to identify optimal modeling parameters for predicting SD. From a ML perspective SD can be modeled and predicted in a generalized manner, by measuring a common measurement like joystick motion or body

movement during labeled moments of SD and non-SD; where labeled moments of non-SD are determined based on correct task completed. During the dataset creation phase, we used existing motion detection experimental design methodologies and designed a generalized motion detection experiment. A vestibular whole-body compensatory task in darkness was used to produce realistic motion cues that a pilot might experience, and motion detection behavior was recorded via joystick movements. Two experiments were conducted: rotational and translational motion detection tasks. The rotational and translational experiments administered angular and linear whole-body stimulation, around and along the three Cartesian coordinate frame axes respectively, using a motion simulation system. Participants were given randomized combinations of three parameters that created angular or linear motion stimuli: axis, direction along the axis, and speed. The goal of the dataset creation phase was to create a realistic and diverse dataset of the perceptual joystick motion with respect to the occurrence of SD. The motivation was not to identify vestibular thresholds and report corresponding behavior, but to recreate realistic flight response data in a controlled manner such that states of disorientation could be modeled. Next, using the generalized SD dataset, ML methods were chosen for SD modeling because of their reliable and effective predictive capabilities [18]. During the comprehensive modeling parameter search phase, we 1) categorized participant responses into four performance categories, 2) created three semi-supervised SD identification labels from the performance categories, 3) created six unique features from the perceptual joystick feedback measure, and 4) compared test set prediction accuracy and receiver operating characteristic area under the curve (ROC-AUC) metrics for five key modeling parameters: feature quantity, model type, dataset conditions, feature type, and semi-supervised label type. ML modeling parameter combinations were identified for accurate SD prediction. The goal of this phase was to create an ML model parameter selection guide for SD prediction such that SD researchers in aviation can readily use successful combinations of parameters with real flight data. Finally, the relationship between physical sickness symptoms and motion disorientation was investigated to identify potential physiological markers for SD prediction; physical sickness was quantified using a generalized disorientation test for humans called the simulator sickness questionnaire (SSQ) disorientation sub-scale [19], [20]. We hypothesized that participants who correctly detected motion, implying that they do not have SD, would not have physical sickness.

II. MOTION DETECTION EXPERIMENTATION

The rotational and translational SD motion detection experiments were designed identically such that the resulting SD dataset would be in a standard format. The following experimental parameters were the same for both experiments: experimental stimulus conditions, number of randomized trials per experiment, timeline of experimental events per trial, experimental protocol, and motion simulation system. In order to validate our SD dataset using our motion simulator platform, we followed recent motion detection protocols that also used motion simulator platforms such as the Moog 6-degree-of-freedom (DOF) motion platform [10], [11], [12], [13]. Recent motion simulator-based motion detection protocols report motion detection thresholds in terms of speed, thus manipulating speed as an experimental condition instead of acceleration. Robotic motion planning is easier and more accurate for speed control than acceleration control. Finally, to create a diverse dataset of vestibular and proprioceptive SD responses, perceptual responses were measured using a 3 x 3 block design testing a randomized combination of angular or linear axis motion, axial direction, and speed. A total of 32 participants, for both experiments, received the same experimental instructions and protocol while using the motion simulator.

A. EXPERIMENTAL DESIGN

The axis experimental condition had three parameters, cabin movements for rotation were roll (RO), pitch (PI), and yaw (YA), and translation included left/right (LR), forward/backward (FB), and up/down (UD). In addition, minuscule sinusoidal vibrational noise of, 1-2 cm in amplitude and frequency greater than 10Hz was added to the non-stimulated axes to mask the sound of the motor for the selected stimulus. Because vibrational noise was present, the participants were exposed to a more realistic aviation environment. Furthermore, the additional vibration helped reduce movement detection thresholds, such that the task was realistically challenging [7]. The axial direction experimental condition had two parameters: positive or negative direction. Figure 1A depicts both the axis and axial direction conventions for both the rotational and translational experiments.

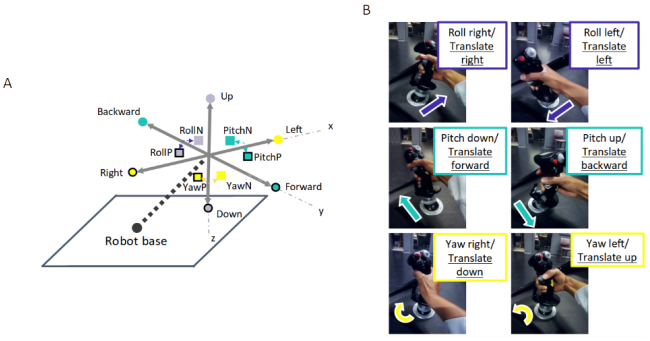

Figure 1. Cabin and joystick axial and/or directional motion convention. The grey Cartesian coordinate frame in Figure 1A represents the simulator cabin. The cabin could move in both rotation (RO, PI, YA) and translation (LR, FB, UD) via the input stimulus and/or participant control. The black outlined squares and circles in Figure 1A denote positive directional movement (RollP, PitchP, YawP, Right, Forward, Down), where squares and circles correspond to rotational and

translation movement respectively. Non-outlined squares and circles indicate negative directional movement (RollN, PitchN, YawN, Left, Backward, Up). Figure 1B shows the mapping of participants' joystick movements to the cabin movement.

The speed experimental condition had two parameters; a slow near below-threshold (sub) speed where motion was difficult to detect and a fast above-threshold (sup) speed where motion detection was easier to detect. Some talented participants could detect sub speed movement, therefore this lower limit perceptual stimulation was emphasized to be at "near below-threshold" instead of at a below-threshold speed. In the motion detection literature, our speed parameters are known as motion detection thresholds measured in terms of Hz, which is a frequency measure of deg/s or cm/s depending on whether the stimulus motion is in rotation or translation respectively. The speed parameters required special selection such that the values would be in alignment with reported motion detection thresholds and accommodate the motion constraints of the simulation system. Rotational and translational, sub and sup speed selection was inspired by reported experimental design thresholds from motion detection literature [11], [10], [12], [13], [9]. A unique list of (0.1, 0.2, 0.5, 0.6, 0.7, 1, 1.25) and (0, 3.75, 5, 7.5, 10, 12.5, 15, 17.5, 20, 22.50) speeds respectively in deg/s and cm/s were constructed for the rotational and translational experiments, respectively. A calibration phase was conducted with 23 naive participants, in which the unique lists of speed values were tested using the motion simulator system. In the calibration phase, participants sat naturally in the simulator and verbally reported the direction in which they believed they were moving directly after randomized motion stimulation; calibration participants were not allowed to participate in the two experiments. Five participants confirmed the rotational calibration experiment, the chosen sub and sup speed that reported the lowest and highest number of correct responses was 0.5 Hz (deg/s) and 1.25 Hz (deg/s) respectively; implying constant acceleration at 0.5 deg/s^2 and 1.25 deg/s^2. Eighteen participants were needed to confirm the translational sub speed value because of variability in response detection across participants was varied, in comparison to the rotational experiment. For the translational experiment, the chosen sub and sup speeds that reported the lowest and highest response accuracy were 3.75 Hz (cm/s) and 15 Hz (cm/s) respectively; implying constant acceleration at 3.75 cm/s^2 and 15 cm/s^2. Specifically, for both experiments, the chosen sup value was the slowest sup speed at which the participant response accuracy was similar to the highest response accuracy.

A single trial was composed of four different phases, as denoted by the timeline in Figure 2, in which participants were tasked to give feedback to specific visual and vestibular stimuli per phase. During phases A and B, participants could move the simulator using the joystick in any of the rotational or translational axes to counteract the perturbation. Joystick control was in terms of velocity control because it allowed for a fast and smooth response signal.

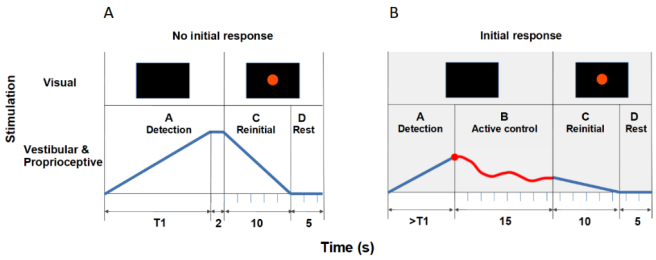

Figure 2. Experimental event timeline examples. Timeline A occurred when the participant did not respond in phase A, it consisted of three phases: (A Detection) motion stimulation of the cabin using a smoothed ramp forcing function, (C Reinitialization) cabin reinitialization to the initial orientation or position, (D Rest) cabin and participant at rest. Timeline B occurred when the participant responded in phase A, the four phases consisted of: A Detection, (B Active control) participant active control, C Reinitialization, D Rest. For both timeline A and B, visual and vestibular stimulation was given during each phase. The blue and red lines are position-based trajectories. The blue line denotes automatic robotic movement of the simulator cabin along one axis per trial, and the red line denotes the stimulus plus the participant's movements to compensate for the perturbation. T1 denotes the maximum allowed stimulation time per trial with respect to each axis and speed, if initial detection was not made within T1s the experimental phases followed as depicted in timeline A. If the joystick was moved within T1s, an initial response was registered and experimental phases occurred as depicted in timeline B.

- Phase A Detection: A smoothed ramp-forcing function, where the rate of displacement was unknown to the participants, slowly and continuously perturbed one of the three rotational or translational axes of the simulator cabin at a sub or sup rate. The acceleration profile was the second derivative of the position trajectory. Position trajectories are shown by the blue and red lines in Figure 2. During phase A participants were tasked to perform "initial detection", which consisted of identifying the axis and direction of the felt perturbation and manipulating a joystick replicating actual aircraft controls (Thrustmaster Hotas Warthog joystick), shown in Figure 1B, in the opposite direction of the stimulus. Participants had 15-20s to detect motion depending on the condition, denoted by T1 in Figure 2, which corresponded to the cabin reaching the maximum allowed cabin displacement. T1 was different for every axis and experiment because sub and sup rates were different for each experiment and the physical cabin displacement range was different for each axis. In particular, the rotational experiment had slightly longer stimulation times than the translational experiment because the sub and sup rates were slower and the available cabin displacements in the RO, PI, and YA orientations were larger than the available translational displacement ranges. If the participants did not respond within T1s during phase A, the cabin automatically displaced along one of the three axes as the ramp function increased until it reached T1s, where the ramp function maintained a zero slope causing the cabin to remain stationary for 2s.
- Phase B Active control: If participants responded within T1s during phase A, phase B active control began and

they had 15s to maintain the simulator orientation or position stably at the initial location by counteracting the perturbation; phase B was a vestibular dead-reckoning task. No visual stimulation was present; thus, the participants could rely only on vestibular and proprioceptive cues.

- Phase C Reinitialization: A red dot appeared on the screen instructing participants to release the joystick and rest, while the cabin automatically returned to the initial starting location within 10s.
- Phase D Rest: The cabin remained stationary at the starting location for 5s in order to avoid possible overstimulation or after-effects.

Figure 2A and 2B show a typical position trajectory when the participant did not respond and when the participant responded during phase A respectively, demonstrating that the experimental phases and trial length were dependent upon the participant's initial response. The shortest and longest trials were approximately 32s and 50s respectively. The shortest trial length occurred when the participant immediately responded within 1-2s (2s+15s+10s+5s) or did not respond with T1 equaling 15s ((15s+2s)+10s+5s), the longest trial length occurred when the participant responded just before T1 with T1 equaling 20s (19.9s+15s+10s+5s).

Both experiments administered 42 trials: 12 familiarization practice trials and 30 experimental trials. During the familiarization practice phase, unique experimental condition combinations were given, where each of the three axes was stimulated in negative or positive directions at sub or sup speeds. Similarly, the experimental phase consisted of 30 randomized trials, in which 15 trials with unique experimental conditions were repeated twice: five direction-speed conditions (negative sup, negative sub, no-movement, positive sup, positive sub) for each of the three axes (RO/LR, PI/FB, YA/UD). No-movement trials were included as sham trials to encourage the participants to remain active.

B. PARTICIPANTS

The EuroMov Institutional Review Board (IRB) at the University of Montpellier approved that the scientific objectives and organization of both experiments (IRB-EM rotational: 1703B, IRB-EM translational: 1704B) were safe and appropriate for human participation. The EuroMov IRB committee rules and regulations are in accordance with the 1964 Declaration of Helsinki and its later amendments. Eighteen and 14 healthy volunteers with normal or corrected vision gave informed consent before participating in the rotational and translational tasks respectively (males and females, 32±10 years old); four of the 32 participants reported having novice time-limited piloting experiences lasting less than 40 hours. Four of the 18 rotational and four of the 14 translational participants were over the age of 40 years. The participants who performed the rotational experiment were not the same than those who performed the translational experiment. Therefore, there was no confounds due to experimental ordering, learning, carryover, or fatigue. The same participant population, university students, and staff, was used for both experiments; therefore, it is likely that both experimental populations were similar.

C. EXPERIMENTAL PROTOCOL AND MOTION SIMULATION SYSTEM

The experiment took approximately 90 min and consisted of four sections (1) arrival, questionnaires, and instruction; (2) familiarization; (3) active control of rotational or translational stimulation; and (4) questionnaire and debriefing. After describing the experimental task and completing the questionnaires, participants were securely installed using the safety harness and communication headphones, as shown in Figure 3A. They were asked to moderately move the joystick in one axis direction at a time while compensating the unknown perturbation. Participants were reminded to maintain the cabin at the initial trial position or orientation, fixed at a steady centered pose, by compensating the motion stimulus. Participants were free to adopt their own strategy to perform the task, both in terms of response speed and of exploration behavior. In order to replicate a realistic flight scenario, the participants were free to move their head and body, looking and/or fixating where they wished, as long as it did not interfere with the task. The fact that the head was left unrestrained is considered undesirable, causing erroneous motion detection due to conflicting self-generated sensory information, and thus rarely performed in traditional motion perception experiments. However we considered it ecologically innovative because it replicated human response under realistic flight circumstances, allowing for a more realistic SD dataset. Once the participant was installed in the cabin, the cabin door was closed and all communication between the participant and experimenter was performed via a camera interface system that facilitated two-way auditory communication. The camera system also provided the experimenter visual feedback of the participant's upper body. The experimenter visually monitored the well-being of the participants, and confirmed participant's feelings of illness auditorily; the experiment ended if the participants reported physical illness.

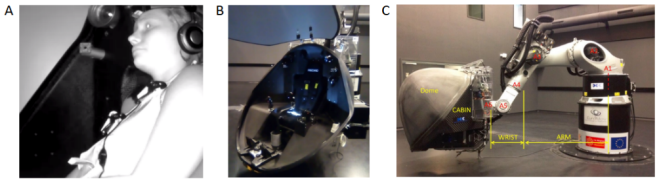

Figure 3. Motion simulator apparatus and installation. A and B show the experimental simulator cabin with and without a seated participant. C shows an exterior view of the six-axis iMose motion simulator, consisting of the participant cabin and the robotic arm.

The motion simulation system that provided sensory stimulation, iMose, consisted of a 6DOF position-controlled KUKA-based motion simulator system (KR 500-3 MT adapted by BEC GmbH motion simulators, KUKA Roboter GmbH, Germany) and a local area network of three

independent workstations [17], [21], [22]. Figures 3B and 3C show the interior and exterior of the simulation system, data was transferred between the simulator and workstations at 250 Hz on a private network using UDP. Workstations 1 and 3 were located in the experimenter control room; workstation 1 generated motion for the robot using a MATLAB/Simulink control interface program (MATLAB and Simulink Toolbox Release 2009, The MathWorks, Inc., Natick, Massachusetts, USA). Workstation 2 was fixed to the simulator cabin, and it administered the red dot or black visual screen and recorded the streamed user-controlled joystick signal. Workstation 3, using Labview, served as the experimenter's user control interface to start and stop the experiment and collect experimental data without causing information delays between the workstations.

Two questionnaires were administered before the experimental phase: a claustrophobia assessment [23], [24] and SSQ [19], [20]. All questionnaires were administered in the native fluently spoken language of each participant (French or English). The claustrophobia questionnaire consisted of two sections: the first section measured fear of suffocation (14 questions) and the second section assessed fear of restriction (12 questions). The claustrophobia questionnaire was used as a screening method to assess whether participants could enter the simulator and perform the task relatively stress-free; participants who scored 40 points or lower, indicating that they were not claustrophobic, were initially recruited, and participants scoring higher than 40 were recruited last. For both the rotational and translational experiments all participants scored "non-claustrophobic", rotational results were mean=10.94, max=38, min=0 and translational results were mean=8.77, max=9, min=0. The SSQ consisted of 16 questions and measured the participant's general physical state, evaluating nausea, ocular motor, and disorientation sub-scales. The SSQ was administered before and after the experiment to measure the effects of the experiment in terms of disorientation.

III. ANALYSIS

As previously mentioned, the goal of this study was to create a realistic flight dataset of moments of disorientation and non-disorientation measured by joystick motion, and then identify the best methods to predict SD using ML methods. The analysis methodologies were as follows: 1) verifying the correctness and authenticity of the dataset, 2) evaluating ML modeling parameters for SD classification using three metrics, and 3) correlating physical with perceptual disorientation to confirm whether other possible measures besides joystick could convey markers for the occurrence of human SD-state. Python was used for all analyses, using numpy, pandas, scipy, scikit-learn, seaborn, plotly, and matplotlib (Python 3.9, Python Software Foundation, Fredericksburg, Virginia, USA).

A. VERIFICATION OF SIMULATION DATASET

All trials, familiarization and experimental trials, were used in data analysis to maximize data usage. The simulator system motion and participant joystick responses were down-sampled from 250 Hz to 10 Hz for data analyses, such that only relevant human motor movements were considered; literature has shown that human hand and arm movements do not exceed frequencies of 10 Hz [25].

Data standardization pre-processing analysis was performed to ensure that the data was collected properly. Data standardization consisted of two-steps: 1) numerical confirmation of experimental settings, and 2) numerical confirmation of the experimental design. In the first step, four items were checked for correctness using both joystick and cabin motion data: experimental event matrix per trial, joystick and cabin directional control convention, joystick margin needed to command cabin motion, and start-stop time of phases A and B per trial. In the second step, the motion and timing of the cabin with respect to joystick motion was checked for correctness. The robotic simulator performed motion stimulation in real time using a real-time Linux kernel, with a MATLAB/Simulink input layer, to capture responses with minimal delay. Despite the advantage of rapid response synchrony, real-time systems are prone to having system delays that can influence functional timing and communication between tasks; real-time functioning refers to the order in which numerical tasks are executed using the available computer resources. Therefore, the rotational and translational experiments had trials where system delays imperceptibly influenced the robotic trajectory and/or the participant's ability to respond correctly via the joystick. Due to these slight processing and thus execution errors that are due to the real-time functionality of the motion simulator, it was necessary to remove all trials that had frequency or joystick-cabin related defects such that experimental defects were not confounded with participant response. The following defects were checked in the second step of data standardization:

- temporal gaps in data,
- trials where phases A and/or B were shorter than the minimal expected trial length of 17s, denoting the system sampling frequency was faster than desired,
- trials where joystick motion was sufficient but the cabin insufficiently moved,
- delays longer than 5s between joystick and cabin movement.

The majority of trials that were removed were due to fast or slow system frequency sampling rates, thus discarding trials that had recorded timestamps less or more than 17 s or 50s respectively. The second reason for discarding trials was due to the fact that the cabin did not respond within a few seconds after joystick movement, or the cabin motion axes and direction was incorrect with respect to joystick motion. In total, 40% of rotational and 50% of translational trial data was removed from the analysis. Errors were expected as the

system was a new experimental test platform where many computers needed to operate in synchrony. Data standardization was the only step that removed trial data, trials that passed data standardization were used in data analysis.

B. RESPONSE CATEGORIZATION

Detection of correct stimuli was categorized into ten possible categories based on the selection of axis and axial direction. Figure 4 depicts a flowchart and possible participant choices based on response movements. The blue squares indicate the experimental trial type: the presence of motion stimuli denoted by "Movement" and no presence of motion stimuli denoted by "Sham". For "Movement" activity, the green squares indicate participant response activity such that "1-3 axes" means that the participant moved the joystick on one or more axes and "No axes" means that the participant did not move the joystick. The yellow diamonds denote the decision process based on the question asked within the diamond. For example, for "Movement" activity where the participant responded using one or more joystick movements, the following question is posed: "Is the stimulus axis the same as the axis in which the participant initially moved the joystick?". If yes, the axis was noted as correct, and the initial direction was confirmed in a similar manner. For example, "Did the participant initially move in the opposite direction of the stimulus direction?". The red numbers indicate the total number of possible categories based on the logical progression of performing the task correctly, first finding the correct axis and then finding the correct direction to counteract the vestibular stimulus.

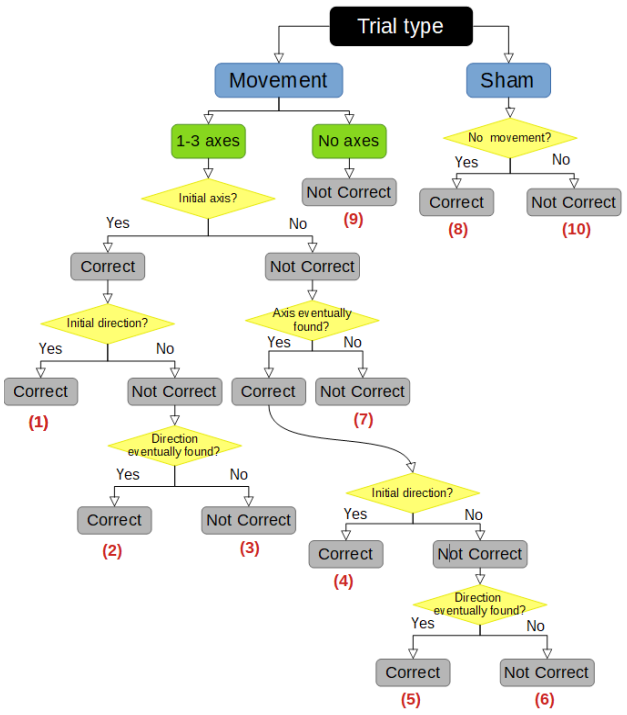

Figure 4. Flowchart of selection process for detection performance categories, where correct response categories 1, 2, 4, and 5 denote non-SD occurrence and wrong response categories 3, 6, 7, and 9 denote SD occurrence.

The ten detection performance categories were reduced to four categories:

- Initially Correct axis and direction: trials in which the first response was with the correct axis and direction (IC: Category 1)
- Eventually Correct axis or direction: trials where the first response was with an incorrect axis or direction but the correct axis and direction was found (EC: Category 2, 4, and 5)
- Never Correct: trials where participants acted on the joystick but never found the correct axis and/or direction (NC: Category 3, 6, and 7)
- No response: trials in which participants did not respond (NR: Category 9).

Categories 8 and 10 corresponded to the no-movement sham trials and were not used in the analysis.

C. MOTION DETECTION AND PERFORMANCE SUMMARY

The normalized response count and Reaction Time (RT) per detection performance category were quantified for each axis and speed condition. The normalized response count was the adjusted count per response category, with respect to the given number of trials multiplied by participants; the total trial count per participant was 36, excluding sham trials. The total trial count per participant was adjusted to 36, such that the interpretation of results would be consistent with the experimental design. Participants had fewer total trials than 36 trials because trials that did not follow the experimental design were removed during the data standardization step mentioned in Section III A. RT was the time that the participant used to find the correct axis and direction. The 95% confidence interval per axis was calculated to determine which detection performance categories were significant. Detection performance categories above the lower confidence interval were evaluated further. Significant and corresponding detection performance categories were compared for the speed and axis.

The Kolmogorov-Smirnov (KS) test was used to evaluate whether to use a parametric or non-parametric two-sample comparison test for within-axis and across-axis comparisons. All test evaluations resulted in non-parametric distributions; therefore, only non-parametric tests were used. Two non-parametric tests were used to evaluate comparisons: Wilcoxon signed-rank distribution test and Wilcoxon rank-sum distribution test [26]. Uneven two-sample non-parametric test data vectors were compared using the Wilcoxon rank-sum test. However, the Wilcoxon signed-rank test required that equal length vectors be compared, thus shorter length vectors were padded with NaN values to preserve the equivalent number of samples with respect to the longer vector and the distribution of the shorter length vector. Statistical p-values are reported using the following

standardized significance levels: the Bonferroni required value of 0.0167 for two test comparisons, 0.05 for single test comparisons, and 0.001 for strongly significant one or two test comparisons.

A participant detection performance rank score was created to compare overall participant detection performance with perfect performance. The performance rank score was calculated per subject across trials, per experiment, where

$$Rank\ score = 2 \cdot (IC\ count) + (EC\ count) \quad (1).$$

The rank score equation weights were arbitrarily chosen such that the equation formulation was most simplistic; RT was not considered in the rank score because rotational and translational experimental stimulation timings were different and thus non-comparable. IC performance was the desired behavior for the task so a weight of two was given to each IC trial. EC was also desired task behavior because participants were able to eventually find the correct axis and direction, however mistakes were made, thus a weight of one was given to each EC trial. NC and NR performance trials were not the desired task behavior so they were given no credit. Thus a rank score of 72 corresponded with perfect performance, where IC detection was performed for all 36 motion stimuli trials. Finally, the rank score was used to divide participants into three final categories in order to summarize performance with respect to each experiment. Mean and standard deviation of participants' rank score per experiment were calculated, such that participants were divided into best, average, and worst categories if their rank score was greater, within, and lower than one standard deviation from the experimental participant mean respectively.

D. CLASSIFICATION MODELS, FEATURE & LABEL CREATION

A total of eight well-known ML techniques were used for SD classification, three decision tree type models and five optimization-based models, in order to evaluate which algorithm structure was most efficient for predicting SD time-series data. Decision tree type models included: Decision Tree (DT), Gradient Boosting Classifier (GBC), Random Forest (RF). The optimization-based models consisted of the following well-known ML algorithm: Stochastic Gradient Descent (SGD), Linear Discriminant Analysis (LDA), 2-Layer Deep Learning Neural Network also called Multilayer Perceptron (MLP), Gaussian Naive Bayes (GNB), Support Vector Machine (NuSVC). Detection classification was performed for each speed condition (sub, sup), and across axis conditions.

The six explored joystick features included:

- joystick signal describing a human movement metric of hand control velocity,
- first derivative of the joystick signal describing a human movement metric of hand control acceleration,
- second derivative of the joystick signal describing a human movement metric of hand control jerk,
- frequency response of joystick signal via the Fast Fourier Transform describing a human-in-the-loop control theory metric,
- joystick signal fundamental frequency calculated by estimating the period of the signal describing a signal processing metric,
- joystick signal fundamental frequency calculated from the bandwidth cutoff or 0.3 decibel drop of the joystick frequency response describing a human-in-the-loop control theory metric.

The fundamental joystick frequency was of interest because it captured lower order harmonics of participant behavior in response to external forces.

Three types of semi-supervised labels were created for predicting disorientation:

- Lenient: mistakes are allowed to be made thus IC and EC categories were labeled as 'non-disoriented' and NC and NR were labeled as disorientation,
- Strict: no mistakes are allowed where IC was labeled as not disorientation and EC, NC, and NR were designated as disoriented,
- Complex: a multi-label depicting SD via NC and NR responses, mild-SD using EC responses, and non-SD using only IC responses.

The purpose of testing different labels was to understand how to best define SD from the intrinsic organization of the data; better predicting models using a certain label implies that the data is best structured for that label. We compare our data-driven definition of SD with the current functional definition of SD [6].

E. CLASSIFICATION MODEL EVALUATION

Average 5-fold cross validation test prediction accuracy and ROC-AUC measures were used to evaluate ML model performance. Accuracy measured the true positive (TP) and true negative (TN) counts over the total number of samples; a value of 1 and 0 correspond to 100% and 0% correct prediction.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

where FP and FN correspond to false positive and negative counts, respectively. Accuracy only gives information about how well the model approves data, but not about how well the model rejects data. Therefore, the familiar ROC-AUC measure was used to evaluate both classification acceptance and rejection performance. ROC-AUC is the area under the false positive rate (FPR), shown in (3), versus the True Positive Rate (TPR), shown in (4),

$$FPR = \frac{FP}{FP + TN}, \quad (3)$$

$$TPR = \frac{TP}{TP + FN}. \quad (4)$$

An ROC-AUC score of one indicates perfect prediction of all labeled classes, whereas a score of 0.5 or lower indicates that prediction of all labeled classes was poor with chance level performance or lower. ROC-AUC was needed in addition to accuracy to determine if FP values were balanced with TP values, ensuring that the SD model could accurately reject and accept the data [18].

Finally, feature importance was of interest because each feature contained distinct information about disorientation. It was of interest to understand which feature/s could convey the most informative information about the occurrence of perceptual disorientation. Feature importance was calculated such that each feature was shuffled individually and model accuracy was calculated for each shuffled feature. Unshuffled model prediction accuracy was subtracted with each of the shuffled feature prediction accuracy scores. The change in prediction accuracy for each shuffled feature was ranked, such that the feature with the largest change in prediction accuracy was considered the most important feature. Feature importance was calculated using scikit-learn's permutation importance function. Individual metric comparisons, of the three metrics, were evaluated using the same Wilcoxon signed-rank or rank-sum tests where $p < 0.05$ and $p < 0.001$ were considered significant and strongly significant respectively; only non-parametric tests were used because the KS test reported non-parametric distributions.

F. PHYSICAL DISORIENTATION

Detection performance categories were related to only the SSQ disorientation sub-scale, not the combined SSQ score, because the task was related to disorientation with respect to motion detection [19], [20]. Physical disorientation was monitored before and after the experiment using the SSQ disorientation sub-scale, such that the difference in before and after measures were attributed to the experienced task; SSQ disorientation difference equaled the disorientation score before the experiment minus the score after the experiment.

Negative SSQ disorientation difference meant that the task made the participant disoriented (e.g., they felt better before), and positive SSQ disorientation difference meant that the task rendered the participant less disoriented (e.g., they felt better after). Physical disorientation for accurate and non-accurate motion detection performers were compared, to quantify whether physical disorientation report could also be a marker for SD, like the perceptual joystick. Again, Wilcoxon signed-rank or rank-sum non-parametric distribution tests were used to evaluate comparisons, as the KS test only found non-parametric distributions. The mentioned statistical p-value reporting convention was used.

IV. RESULTS

For both rotational and translational experiments, participants' detection behavior was quantified using the detection performance categories in terms of count and initial RT with respect to stimulation speed and axis; no significant differences were found between positive and negative axial directions thus directional differences were not considered. As previously mentioned, human motion detection ability of self-motion along axis and axial directions are often reported in terms of motion detection threshold [13], [11], [12]. A motion detection threshold is registered from a self-report that motion was felt along a specific axis and direction for a specific motion stimulus frequency. Results are typically displayed in terms of mean detection count across or per subject for many stimulus motion frequencies, where count results are grouped by successful and unsuccessful detection. We performed the same analysis presentation for our two sub and sup stimulus frequencies, displaying results in terms of count, mean count, and RT across participants, where count and RT results are grouped by detection performance categories.

A. MOTION DETECTION PERFORMANCE

Figure 5 shows the normalized summed count (top row), normalized mean count (middle row), and mean RT (bottom row) per detection performance category, across participants for axes and speed conditions for both rotation and translation. The top row shows the normalized summed count per response category for each axis and speed condition. For rotation, the summed bars in the top row are equal to 648, which corresponds to the 18 participants multiplied by 36 trials. The top row represents the distribution of trial responses per response category. Similarly, for translation, the summed bars in the top row are equal to 504 which corresponds to the 14 participants multiplied by 36 trials. The middle row shows a different visualization of the same normalized count data, except it displays the mean count. The mean count represents the frequency of selecting a response category across participants. Lastly, the bottom row displays the mean RT taken to detect correctly, thus only IC and EC response categories are shown. Bars without error bars indicate a single sample value, or several participants had the same count value. Single-sample bar values may exist due to the rigorous standardization process.

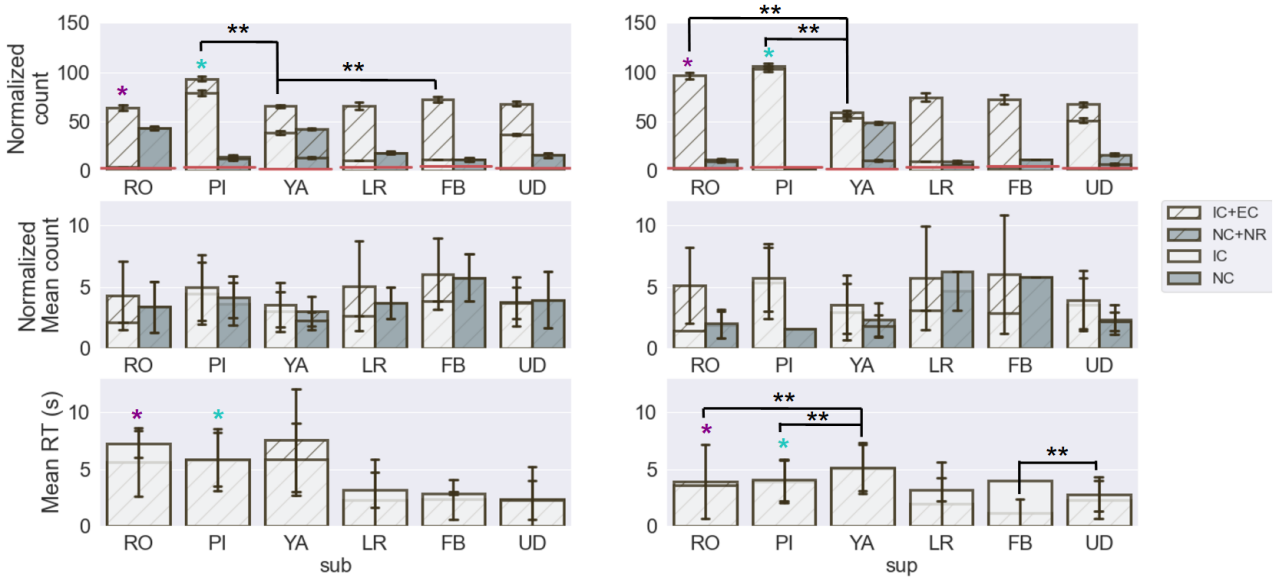

Figure 5. Normalized summed count (top row), normalized mean count (middle), and mean RT in seconds (bottom row) per detection performance category for rotational and translational stimulation. The three metrics were compared for sub and sup speed stimulations for RO, PI, YA, LR, FB, and UD axes respectively; Wilcoxon signed-rank test and rank-sum tests were used to determine significance such that significant and slightly significant relationships were represented by (* within axis comparison of sub and sup, ** across axes comparison). Bonferroni correction: $p < 0.0167$ was used as the significance threshold. Detection performance categories above the lower confidence interval, denoted by the solid red line, were considered for statistical comparison across subjects for categories within (e.g.; sub vs. sup) and across axis (e.g.; RO sub vs. PI sub) conditions.

1) DETECTION: SPEED COMPARISON (WITHIN AXIS CONDITION)

There was a slightly significant sub versus sup count difference for the most counted detection performance category for the RO and PI axes, where sup speed resulted in a higher count than sub speed (RO count EC sup vs sub: KS: non-normal distribution, signed-rank: $p < 0.001$, rank-sum: $p < 0.026$, n=13; PI count IC sup vs sub: KS: non-normal distribution, signed-rank: $p < 0.08$, n=18). Slight significance is denoted in the top row of Figure 5, with a single star in purple and blue. There was a similar trend for the YA axis, where the most counted detection performance category, IC, had a higher sup count than sub count. This demonstrates that participants were more accurate at sup speed than at sub speed, regardless of the motion stimulation axis or direction. No significant differences between sub and sup speeds were found in the translational experiment. However, there was a trend for all axes where the most counted detection performance category had higher sup counts in comparison to sub counts. For example, the EC detection performance category for the LR and FB axes and the IC detection performance category for the UD axis had greater sup counts than sub counts. Translational motion sub and sup speed differences were less apparent than in rotational motion due to inner-ear stimulation differences. Reduced speed detection in translational motion are likely attributed to less semi-circular stimulation and delayed otholic signaling in comparison to rotational motion [8]. Therefore, we suspect that more data was needed for differences to become statistically significant.

Regarding RT differences for the rotational experiment, some detection performance categories had significantly lower RTs for the sup than the sub speed condition. For RO and PI axes, the RT for the most counted detection performance category was significantly lower for sup speed in comparison to sub speed (RO RT EC sup vs sub: KS: non-normal distribution, signed-rank: $p < 0.001$, rank-sum: $p < 0.001$, n=66; PI RT IC sup vs sub: KS: non-normal distribution, signed-rank: $p < 0.001$, rank-sum: $p < 0.001$, n=65); significance is denoted in the bottom row of Figure 5 with a single star in purple and blue. In summary, we demonstrate that faster sup motion caused more accurate and faster motion detection than slower sub motion. This result has been reported in motion detection literature, thus confirming that the experiments were performed correctly and that the dataset accurately represented human response [13], [11].

2) DETECTION: AXES COMPARISON (ACROSS AXES PER SPEED CONDITION)

The same speed condition and successful response categories denoted by IC and EC were compared across axes, in order to identify task difficulty with respect to the axis, and thus demonstrate that our experimental results were in alignment with psychophysical motion detection findings. Whole-body motion detection literature shows that RO and PI are easier to detect than YA and translational motion. According to a report RO and PI detection thresholds were statistically similar for novices and experts [11]. Additionally, RO was reported to be easier to detect than LR, UD, and YA in both non-vestibular and vestibular dysfunction participants [13]. Table I shows that, in alignment with literature reports, we similarly find that RO, PI, and FB axial motions were easier to detect than YA, with dependence on speed when considering only correct responses. In particular, successful

response category counts for both RO and PI at sup speed were significantly higher than those for YA, and for sub speed FB and PI had significantly higher counts than YA.

TABLE I
COUNT AND RT COMPARISONS FOR COMBINED IC & EC RESPONSE

	Speed & axis category		Significance (KS: non-normal, signed-rank, rank-sum)
	Category 1	Category 2	
Counts	sup RO	sup YA	$p < 0.001$,0.0167, n=20
	sup PI	sup YA	$p < 0.001$, 0.0167, n=20
	sub FB	sub YA	$p < 0.001$, 0.0167, n=22
	sub PI	sub YA	$p < 0.001$, 0.001, n=22
RT	sup RO	sup YA	$p < 0.001$, 0.001, n=67
	sup PI	sup YA	$p < 0.001$, 0.001, n=67
	sup FB	sup UD	$p < 0.001$, 0.001, n=41

Response count and mean RT speed and axis category comparisons, in order to rank ease of axis and speed motion detection with respect to literature reports; the first and second p-values correspond to the signed-rank and rank-sum test respectively. Category 1 and 2 represent categories with high count or fast RT and categories with low count or slow RT respectively.

Moreover, our results showed functional differences between the RO, LR, & FB and PI, YA, & UD tasks. During PI, participants mostly detected correctly (IC) and rarely when they did not detect correctly, they eventually or never corrected. Again in YA, participants often detected correctly (IC), but when they did not detect correctly they did not feel any motion (NR). Similarly in UD participants often detected correctly (IC), and when they did not detect correctly they often eventually corrected; IC was more prevalent when speed was fast. Whereas in RO, LR, and FB, participants could not initially detect the correct axis and/or direction, but they could eventually find the correct axis after several mistakes.

Lastly, task difficulty for within rotational and translational stimulation appeared to be correlated with longer RT. As mentioned in Section II A, participants were stimulated slower in the rotational task than in the translational task; thus, RT was different for the rotational and translation tasks and were not compared. The second portion in Table I labeled RT, shows the significant within experiment comparison across axes for only correct responses. For the rotational task at sup speed, RO and PI had faster RT than YA indicating that participants needed less time to detect motion for RO and PI. Similarly, for the translational task at sup speed, FB had significantly faster RT than UD.

B. MOTION DETECTION PERFORMANCE RANK

Including both rotational and translational tasks, the highest rank score was 55 and the lowest score was 11. On average, participants received a rank score of 37. Therefore, the best performer, regardless of rotation or translation, achieved $(55/72)\cdot100=76.38\%$ accuracy for the task. The average performer was only able to achieve $(37/72)\cdot100=51.38\%$ accuracy for the task. The same task accuracy statistic was calculated for sub and sup conditions individually, for both rotation and translation experiments, and similar results were found, as shown in Table II.

TABLE II
DETECTION PERFORMANCE RANK PER SPEED CONDITION

Rank	Rot sub	Trans sub	Rot sup	Trans sup
Best	83.3%	75%	80.5%	77.7%
Average	47.2%	55.5%	55.5%	58.3%
Worst	11.1%	0%	30.5%	25%

Experimental performance accuracy per speed condition using the performance rank measure. All percentages were calculated by dividing by 36 trials.

These rank statistics show that the detection task was challenging for the average person, regardless of the experimental conditions, but it was not impossible to perform with reasonable success. The participant distribution count for the rotational experiment was five best performers, 11 average performers, and two worst performers. Similarly, the participant distribution count for the translational task was as follows: two best performers, 11 average performers, and one worst performer. The rotational and translational participant distribution counts for best, average, and worst performance were similar, showing that both tasks were similarly challenging in terms of motion detection. Therefore, translational detection may not be more difficult than rotational detection in realistic environments.

C. SD CLASSIFICATION

Classification accuracy was used to evaluate the predictive ability of each of the eight models. These accuracy results not only directly tell us which models are superior to others for given parameters, but also indirectly tell us about the data. Figure 6 shows that decision tree type models (DT, GBC, RF) were superior to optimization-based models (SGD, LDA, MLP, GNB, NuSVC) using the given features, regardless of the experiment type (rot, trans), axis, and speed. Indirectly, this implies that important associations for SD prediction are not related to temporal or ordered information, because decision-tree-based models cannot capture temporal patterns well. Additionally, we trained models and tested both training and test hold-out data for both originally ordered features and reordered features with a normal distribution, accuracy results were similar for both analyses; normal distribution results are shown in Figure 6. The fact that the prediction results were similar demonstrates that the

temporal ordering of movements was not a strong indicator of SD but rather the amplitude values.

By comparing the three different semi-supervised labels, it was shown that better accuracy can be obtained for optimization-based models using a binary label in comparison to a multi-label (averaged lenient and strict vs complex: KS: non-normal distribution, rank-sum: $p < 0.05$, n=5). In particular, the lenient label for the translational task produced higher accuracy for optimization-based models (trans lenient vs complex: KS: non-normal distribution, rank-sum: $p < 0.05$, n=5). However, regardless of the labels, there was no difference in the accuracy of the decision-tree-based models. Indirectly, regarding optimization-based models, this again indicated that the SD and non-SD responses were similar. Labeling with respect to correctness, the lenient label, better separated data with respect to the amplitude than labeling based on the initial correctness.

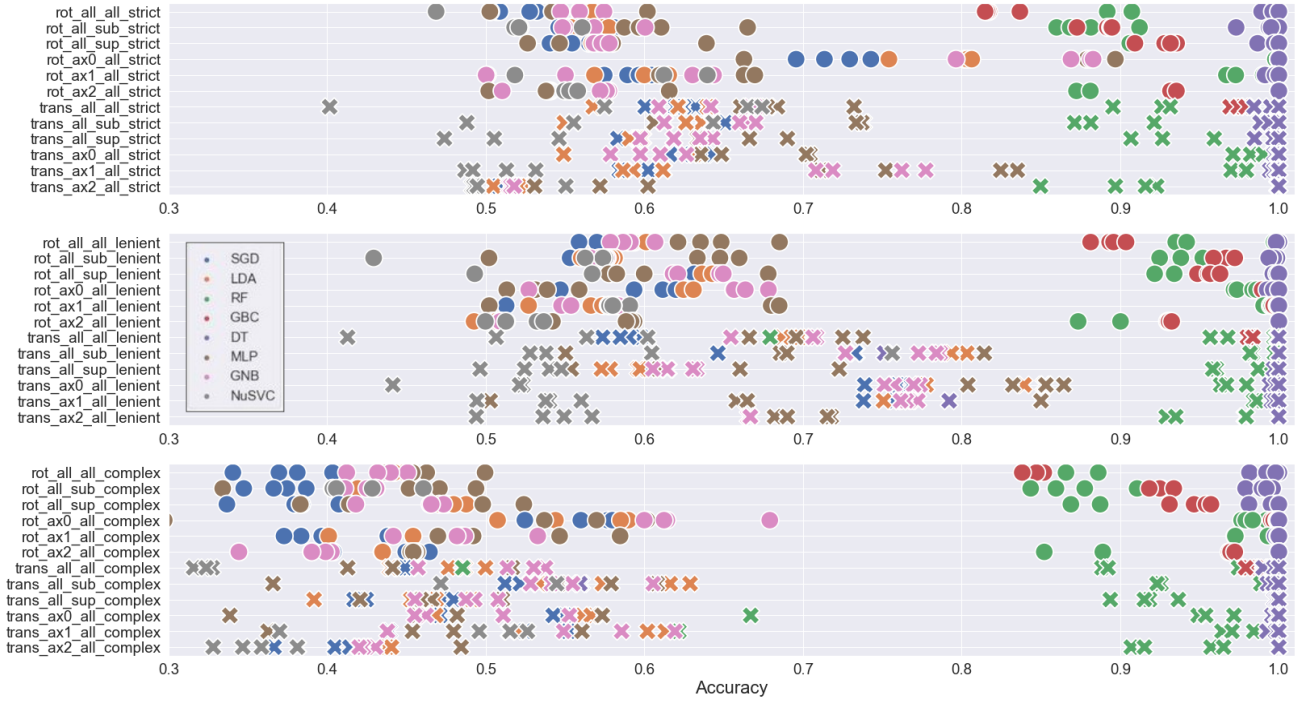

Figure 6. Average 5-fold cross validation test prediction accuracy for the eight models for different: experimental data subsections, semi-supervised label (strict, lenient, complex), and number of features used. For visual ease, circles and crosses correspond to the rotation and translation experiment types, respectively. Blue, orange, green, red, purple, brown, magenta, and gray correspond to the SGD, LDA, RF, GBC, DT, MLP, GNB, and NuSVC models. Regarding notation, ax0, ax1, ax2 indicate RO, PI, YA for the rotational task and LR, FB, UD for the translational task. Additionally, sup, sub, and all denote sup speed stimulation trials, sub speed stimulation trials, and all trials, respectively.

Finally, we investigated whether modeling a subset of the data, such as an experimental axis or speed condition, would result in better prediction of SD for its use-case in comparison to predicting with a model built from combined use-cases. Subset data models were trained and tested using data from specific experimental conditions, such as the RO axis for both sub and sup speeds. Subset data models were not tested on different experimental condition groups or use-cases, because model transferability was not of interest. It was of interest to understand if specific use-cases needed their own individual models, or whether modeling combined use-cases was sufficient for predicting across use-cases, on average. Three dataset model test-prediction accuracy groups were compared using a non-parametric distribution test for each of the three labels: 1) test prediction accuracy of models built from all the data, 2) averaged test prediction accuracy of models built from specific axis data, and 3) averaged test prediction accuracy of models built from specific speed data. No significant differences were found with regard to using all of the data in comparison to only speed or axis stimulus data, for decision tree models. Optimization-based models had four cases in which there were accuracy differences with respect to the subset of data used. Rotation axis data subset had higher accuracy for strict and complex than using all of the data (rot axis strict vs all data: KS: non-normal distribution, signed-rank: $p < 0.05$, n=8; rot axis complex vs all data: KS: non-normal distribution, signed-rank: $p < 0.05$, n=8). There appears to be a predictive effect when using an initially correct label convention, strict or complex label, for axis condition data. Additionally, translational axis and speed data subset for the complex label had higher accuracy than using all of the data (trans speed complex vs all data: KS: non-normal distribution, signed-rank: $p < 0.05$, n=8 ; trans axis complex vs all data: KS: non-normal distribution, signed-rank: $p < 0.05$, n=8). It is likely that using subset speed and axis data, instead of all of the data, helped to reduce variability such that distinguishing patterns could be found.

The ROC-AUC measure was used to evaluate the models' predictive ability with respect to the label value, to evaluate whether the models could predict correctly regardless of the label value. The ROC-AUC parameter combination results were similar to the accuracy results, indicating that the models can reliably detect both SD and non-SD classes; thus, the accuracy results based on TP and TN counts can be trusted.

ROC-AUC main results were:

1. Decision tree type models (DT, GBC, RF) had almost perfect prediction regardless of the semi-supervised label and optimization-based models (SGD, LDA, MLP, GNB, NuSVC) with binary labels had significantly better prediction accuracy than the complex multi-label (averaged lenient and strict vs complex: KS: non-normal distribution, rank-sum: $p < 0.05$, n=5).
2. Moreover, of the two binary labels for the translational task, the lenient label produced more accurate prediction models than the complex multi-label (trans lenient vs complex: KS: non-normal distribution, rank-sum: $p < 0.05$, n=5).
3. Models using all the data instead of a subset of the data, provided similarly accurate performing models except for the same four cases mentioned above for the accuracy results.

Feature order importance allows us to determine which aspect of the joystick signal contains key information to detect spatial disorientation.

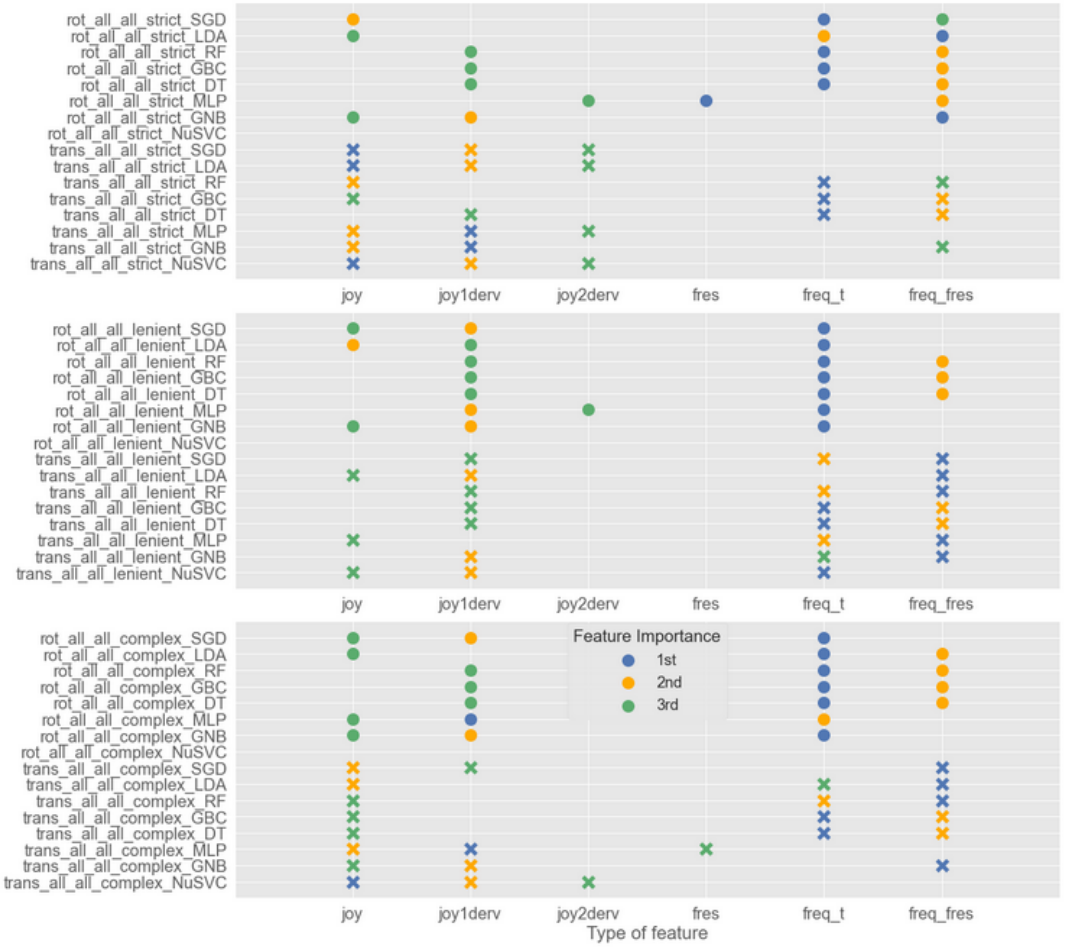

Figure 7. Feature order importance based on permutation importance, for each model. Blue, orange and green circles indicate most important feature, second most important feature, and third most important feature.

Figure 7 shows the feature importance for the models in which all rotational or translational data were used. Three main findings were identified: 1) decision tree models perform optimally using extracted temporal information constant features, 2) optimization-based model feature importance identified more temporal features as being more important than constant features, and 3) semi-supervised labels did not influence feature importance for decision tree models however the strict label for optimization-based models identified only temporal features as being important.

Extracted temporal information, such as the natural frequency of the signal that was transformed into a constant scalar feature, greatly improved the prediction accuracy of the decision tree models. Regardless of the semi-supervised label type and experiment, decision tree models relied on natural frequency-based features more than position based

features, to predict SD. It is likely that these models found stronger patterns with natural frequency features because these feature values repeat per trial, instead of having a unique temporal trace like position based and the frequency response features. Decision tree model prediction accuracy was tested only using the four temporal features, and accuracy results were substantially less accurate at an average of 0.7 in comparison to 0.9 when the two natural frequency constant features were included. Decision tree models are good at finding recurring/mode-like patterns between the features and the corresponding labels, whereas the optimization models are good at finding recurring patterns coupled with specific constraints such as causality. In general, feature importance for decision tree models, from most to least important, were: 1) constant natural frequency of the joystick signal using a periodic estimation method, 2) constant natural frequency of the joystick signal using a frequency response method, and 3) the derivative of the joystick signal. Optimization-based models were shown to be more sensitive to temporal features than decision tree models, where more temporal features were identified as more important than constant natural frequency features. On average, feature importance for optimization-based models, from most to least important, were: 1) constant natural frequency of the joystick signal using periodic estimation, 2) the derivative of the joystick signal, and 3) the joystick signal.

D. MOTION DETECTION PERFORMANCE AND PHYSICAL DISORIENTATION

Twenty of the 31 participants did not feel any difference in terms of physical disorientation during the entire experiment. Considering the performance rank score mentioned in Section III C, approximately 1/3 of the average detectors, 1/3 of the best detectors, and 2/3 of the worst detectors experienced physical disorientation. The 2/3 worst detection ratio is reported for completeness; however this measure is disregarded because it is based on only three participants. Thus, 1/3 of the population felt physical disorientation regardless of performance.

To investigate whether there was a relationship between physical disorientation and detection performance, the detection performance of the 12 participants who reported physical disorientation was evaluated; see Table III for a percentage of their summed trial performance per category per SSQ difference report. For instance, a participant who reported a before and after SSQ score of six and four respectively would have their trial performance category counts, of eight EC and six IC trials, associated with an SSQ disorientation difference score of negative two. Table III demonstrates that more negative physical disorientation was observed for unsuccessful initial detection response categories EC and NC, than for IC successful initial detection response or no response. The negative and positive SSQ disorientation differences per response category were summed respectively, to evaluate significance between IC negative and NC or EC negative. Physically disoriented best performers (IC) did not report significantly less physical disorientation than poor performers.

TABLE III
SSQ DISORIENTATION SUB-SCALE PER MOTION DETECTION PERFORMANCE CATEGORY

	SSQ (%)						
Category	-5	-3	-2	-1	1	2	4
IC (I)	5.4	8.4	17.8	23.6	24.5	8.9	11.5
EC (2,4,5)	**10.1**	**12.0**	**29.0**	**26.6**	14.1	2.8	5.4
NC (3,6,7)	**10.3**	**2.7**	**38.5**	**26.3**	5.2	11.5	5.5
NR (9)	0	3.4	20.8	18.7	26.0	21.8	9.4

Motion detection response category per reported SSQ disorientation sub-scale difference for both the rotational and translational experiments. Performance category percentage values across SSQ scores sum to 100%. Negative and positive SSQ values denote that the participant felt better before and after the task respectively. The bold percentages corresponding to negative SSQ values for categories EC and NC highlight that more negative physical disorientation was present in unsuccessful initial attempts to detect motion.

In summary, we found no significant relationship between physical disorientation and motion detection. There was only a trend that EC and NC performers, who felt physical disorientation, felt better before the task than after. This implies that participants became fatigued while trying to perform the task, when detection was not easy for them. For IC performers who experienced physical disorientation, there was no trend in terms of feeling better before or after. Implying that participants who could detect easily, felt discomfort for other reasons not related to the experiment. There was a slight trend for NR performers that felt physical disorientation, such that they felt better after the task than before. Showing that participants who did not respond, became comfortable and relaxed in the dark experimental setting.

V. CONCLUSION

In this comprehensive study on SD, we show that it is possible to isolate, simulate, and recreate realistic aspects of a vestibular feedback dead-reckoning piloting task and predict the occurrence of SD using joystick response as a feature. We demonstrate that SD can be modeled in a generalized manner with respect to task performance (e.g., correct, not correct) and a human behavioral measure which was joystick motion. Using the ML framework of label and feature organization, SD can be predicted accurately for each SD use-case using use-case specific joystick data or non-specific use-case joystick data. Additional, if other human behavioral measures were included as features such as physiological measures, ML classification could include both prediction of SD and prediction of specific use-case. Thus,

SD can also be studied and predicted using a data-driven ML task-measurement approach, instead of the widely practiced functional use-case driven approach.

The generalized experimental design allowed for the collection of perceptual response joystick data during various basic scenarios of vestibular and proprioceptive stimulation; SD or non-SD states were apparent via the joystick response. A crucial data standardization step was used to verify that the simulator system correctly performed the experimental design, removing trials with delays and erroneous motion. The SD-targeted dataset captured known human motion detection trends, demonstrating that the real-time motion simulation environment was fidel despite the functional timing delays [27]. To mitigate functional timing issues, where some events were executed incorrectly before other events, programming events could have been grouped into functional blocks or scripted codes, where similar tasks were executed in synchrony. Known motion detection trends include: a) accurate and faster response for sup speed stimulation in comparison to near below-threshold speed stimulation, b) PI, RO, FB, LR, UD, and YA were the least to most difficult axis tasks, c) longer reaction times corresponded with task difficulty for the respective rotational and translational experiments [13], [11], [12]. Ranking of task difficulty per axis confirms literature reports that there is no sensory advantage for UD detection due to gravity, because the vestibular system compensates for gravity [13]. In addition to confirming known motion detection trends, functional differences in motion detection for RO, LR, & FB and PI, YA, & UD tasks were observed; where the most counted response for RO, LR, & FB was EC in comparison to IC for PI, YA, & UD. It is unclear why participants made more initial mistakes for RO, LR, & FB than PI, YA, & UD axes, however perhaps participants relied on more non-vestibular sensory cues (proprioception, tactile, auditory from the simulator motor) and/or had better natural upright posture during certain stimulus motions than others. Perhaps in PI, YA, and UD they relied more on clear vestibular cues because they self-generated less additional motion information from self-motion or joystick interaction, thus allowing participants to either initially detect correctly or not. It appears plausible that in RO, LR, and FB participants were more likely to generate additional and perhaps conflicting sensory information by naturally tilting or turning the head. It is likely that participants naturally adapted their posture, to be more or less upright, during certain motion stimuli in comparison to others. For example, a slightly left tilted head during pitch motion would more likely induce discomfort than during roll motion, thus encouraging participants to naturally sit upright during pitch stimuli and thus giving them an advantage to detect the motion more clearly. Such functional errors in RO, LR, and FB caused by natural postural behavior could easily escalate the occurrence of spatial disorientation.

Statistical analysis showed that regardless of experimental conditions, the best performers achieved 76% detection accuracy and average performers achieved 51% accuracy. The task may have been difficult because participants were given the freedom to decide on which axis and in what direction the stimulation occurred, as is done in a real-life piloting situation. All participants had very little to no piloting experience, thus our results reflect human motion perception without the influence of piloting experience or training. Thus, the modeling results obtained from this dataset may not be representative of expert piloting behavior, because our novice participants' responses have more variability than expert piloting responses.

Using the SD dataset we investigated modeling methods for predicting SD, including statistical analysis, predictive control, and ML. Predictive control has been used to predict motion detection [14]; however, ML methods have been shown to be efficient and accurate in finding patterns among different types of data and constructing reusable models for future prediction [18]. Using ML techniques, we investigated parameter tuning selection for the prediction of SD and explained the importance of different parameter selection methods. We evaluated the importance of model construction parameters using test set prediction accuracy and ROC-AUC as a benchmark and comparative measure. Five key model construction parameters were tested: feature quantity, eight model types, dataset conditions, feature type, and semi-supervised label type.

Regarding the number of features in each ML model, no significant difference was found in the prediction accuracy when using all six joystick features, three of the most important features, two of the most important features, or the most important feature alone. Thus, building a model on a single important feature is sufficient to predict SD. However, it is a best practice to use all relevant features for model construction. Decision tree type models (DT, GBC, RF) were superior, to optimization-based models (SGD, LDA, MLP, GNB, NuSVC) using simplistic features, in test accuracy prediction regardless of the experiment (rot, trans), axis, speed, and semi-supervised label. On average decision tree models had accuracy rates ranging from 0.8-0.99 depending on the model type. These models were able to learn associations between features and labels regardless of semi-supervised label construction. However, optimization-based models predicted the best when labels were constructed with a binary label instead of a multi-label. Specifically, the lenient label resulted in better prediction than the strict label. On average optimization-based models with binary labels had accuracy rates ranging from 0.5-0.85. Moreover, specialized models for axis or speed conditions did not outperform models in which all data were used for decision tree models. Whereas, for optimization-based models, some specialized models predicted better than models in which all the data was used, implying that lower data variability is important for SD prediction using

optimization-based models.

Three main findings were identified during feature importance investigation: 1) decision tree models perform best using constant natural frequency features, 2) optimization-based model feature importance identified more temporal features as being important than constant natural frequency features, and 3) semi-supervised labels did not influence feature importance for decision tree models however the strict label for optimization-based models identified only temporal features as being important. When a ML model performs poorly or does not select a feature as being important, it means that the optimization strategy cannot produce a prediction in alignment with the label, using the given feature information. Optimization-based models poorly use features with repeating constant values because the optimization strategies often require feature data point distances to be optimized in a certain manner. Feature data points with the same value do not allow for the optimization algorithms to find optimal maximization or minimization predictions. In hindsight, instead of using constant natural frequency features as a simple categorical-type feature, it would have been more insightful for optimization methods to create a more complex natural frequency feature using wavelets or PCA, and/or features created from a clustering method like kmeans. Despite the fact that decision tree models appear to be more accurate and easier to tune than optimization-based models, optimization-based models are useful for data with stationarity and merit the additional time to optimally tune these models. An optimally-tuned optimization-based model would be able to predict the occurrence of SD while accounting for stationarity trends, like those that occur during the leans SD use-case, while decision tree models are more likely to confuse the trend with SD related behavior.

Finally, the decision to choose a strict or complex label convention versus a lenient label to discern SD depends on the application and the quantity of data available to represent a non-SD state. The strict label was less realistically representative of non-SD, because perfect behavioral data in any task is statistically rare, thus building an accurate model using a strict label convention maybe more challenging. Similarly, for the complex semi-supervised label there was a lack of representative data for each of the SD cases. Over-fitting, where training predictions were higher than test predictions, was observed for both strict and complex labels supporting a lack of data for these labeling conventions. The lenient label, labeling with respect to overall correctness regardless if mistakes are made, was shown to be the best labeling convention as: 1) over-fitting was less observed for our small dataset where test and training predictions were similar, 2) both natural frequency and temporal features were selected as important features thus a diversity of relevant simple patterns were captured. Moreover, prediction accuracy with respect to semi-supervised label construction can be used as a confirmation for how to define SD because semi-supervised means that the label is not 100% the ground truth. The prediction accuracy of the model construct using the data, conveys missing information about the label based on trends in the data. If a model predicts better with label A than label B it means that label A better matched the existing data structure, thus confirming that label A is likely to be the ideal label for the data. Therefore, applying this idea to the three semi-supervised SD labels. The lenient label is the ideal label convention for modeling SD, with respect to our dataset, because on average both decision tree and optimization-based models predicted the test data best using the lenient label. A lenient label convention, where SD is defined as never correct or no response, is also in alignment with the definition of SD, where SD is defined as involving successive failures and major performance errors [6].

One-third of the participants experienced physical disorientation during the task; however, no significant relationship between physical disorientation and motion detection was found. There was a trend where participants who initially detected unsuccessfully felt worse after the experiment than participants who did initially detected successfully or did not try. More sample points regarding physical disorientation are needed during the experiment, instead of a sample before and after the task, in order to determine if physical disorientation is correlated with motion detection. In summary based on our correlation results two-sample before and after SSQ questionnaire methods, that measured sickness caused by inner-ear stimulation, were insufficient to uncover correlations with perceptual disorientation. We do not claim that questionnaire methods can not quantify SD, however before and after questionnaire samples may not produce enough data to find statistically significant correlations with other SD measures especially when population sample size is small. A physiological sampling measure that implies physical discomfort, with a sampling rate comparable to that of the joystick, such as EEG, NIRS, heart rate, or electrodermal activity, could provide more insight into correlations with physical and perceptual disorientation.

REFERENCES

[1] K. K. Gillingham and F. H. Previc, “Spatial orientation in flight,” Armstrong Laboratory, Brooks Air Force Base, TX, USA. Rep. no., AL-TR-1993-0022, 1993.

[2] W. Bles, “Spatial disorientation training demonstration and avoidance,” North Atlantic Treaty Organisation (NATO), Soesterberg, Netherlands. Rep. no., TR-HFM-118, 2008.

[3] R. Gibb, R. Gray, and L. Scharff, “Spatial disorientation—cues, illusions and misperceptions,” in *Aviation visual perception: Research, misperception and mishaps*. Farnham, Surrey, UK and Burlington, VT, USA: Ashgate, 2010.

[4] G. Perdriel and A. J. Benson, “Spatial disorientation in flight: current problems,” Advisory Group for Aerospace Research and Development, Neuilly-sur-Seine, France. Rep. no., AGARD-CP-287, 1980.

[5] F. H. Previc and W. R. Ercoline, *Spatial disorientation in aviation,* Reston,VA, USA: American Institute of Aeronautics and Astronautics, 2004.

[6] D. G. Newman and A. FAICD, "An overview of spatial disorientation as a factor in aviation accidents and incidents," Australian Transport Safety Bureau, Canberra City, Australia. Rep. no., B2007/0063, 2007.
[7] S. E. Chaudhuri, F. Karmali, and D. M. Merfeld, "Whole body motion-detection tasks can yield much lower thresholds than direction recognition tasks: implications for the role of vibration," *J Neurophysiol*, vol. 110, no. 12, pp. 2764-2772, 2013.
[8] D. E. Angelaki and K. E. Cullen, "Vestibular system: the many facets of a multimodal sense," *Annu Rev Neurosci*, vol. 31, no. 1, pp. 125-150, 2008.
[9] J. G. Melvill and L. R. Young, "Subjective detection of vertical acceleration: a velocity-dependent response," *Acta Otolaryngol*, vol. 85, p. 45-53, 1978.
[10] M. C. Bermúdez-Rey, T. K. Clark, W. Wang, T. Leeder, Y. Bian, and D. M. Merfeld, "Vestibular perceptual thresholds increase above the age of 40," *Frontiers in Neurology*, vol. 7, p. 162, 2016.
[11] M. Hartmann, K. Haller, I. Moser, E.-J. Hossner, and F. W. Mast, "Direction detection thresholds of passive self-motion in artistic gymnasts," *Exp Brain Res*, vol. 232, no. 4, pp. 1249–1258, 2014.
[12] F. Karmali, M. C. Bermúdez Rey, T. K. Clark, W. Wang, and D. M. Merfeld, "Multivariate analyses of balance test performance, vestibular thresholds, and age," *Frontiers in Neurology*, vol. 8, p. 578, 2017.
[13] Y. Valko, R. F. Lewis, A. J. Priesol, and D. M. Merfeld, "Vestibular labyrinth contributions to human whole-body motion discrimination," *J Neurosci*, vol. 32, no. 39, pp. 13537–13542, 2012.
[14] F. Soyka, P. R. Giordano, K. Beykirch, and H. H. Bülthoff, "Predicting direction detection thresholds for arbitrary translational acceleration profiles in the horizontal plane," *Exp Brain Res*, vol. 209, no. 1, pp. 95–107, 2011.
[15] B. Cheung, K. Hofer, C. J. Brooks, and P. Gibbs, "Underwater disorientation as induced by two helicopter ditching devices," *Aviation, Space, and Environmental Medicine*, vol. 71, no. 9, pp. 879–888, 2000.
[16] J. Sargent, S. Dopkins, J. Philbeck, and J. Arthur, "Exploring the process of progressive disorientation," *Acta Psychol*, vol. 129, no. 2, pp. 234–242, 2008.
[17] F. Denquin, J. Foucher, S. Pla, J.-C. Sarrazin, and B. G. Bardy, "Optical and gravito-inertial contributions to the perception and control of height in a simulated low-altitude flight context," *Ergonomics*, vol. 64, no. 10, pp. 1297–1309, 2021.
[18] A. Burkov, *The Hundred-Page Machine Learning Book*. Andriy Burkov Canada, 2019.
[19] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *International Journal of Aviation Psychology*, vol. 3, no. 3, pp. 203–220, 1993.
[20] S. Bouchard, G. Robillard, and P. Renaud, "Revising the factor structure of the simulator sickness questionnaire," *Annual Review of CyberTherapy and Telemedicine*, vol. 5, pp. 117–122, 2007.
[21] J. Landrieu, J. Abdur-Rahim, J.-C. Sarrazin, and B. Bardy, "Time-to-collision estimates during congruent visuo-vestibular stimulations," in *Studies in Perception and Action XIV: Nineteenth International Conference on Perception and Action (IPCA)*, pp. 109–112, Psychology Press, 2017.
[22] T. Bellmann, J. Heindl, M. Hellerer, R. Kuchar, K. Sharma, and G. Hirzinger, "The dlr robot motion simulator part i: Design and setup," in *2011 IEEE International Conference on Robotics and Automation*, pp. 4694–4701, IEEE, 2011.
[23] A. S. Radomsky, S. Rachman, D. S. Thordarson, H. K. McIsaac, and B. A. Teachman, "The claustrophobia questionnaire," *Journal of Anxiety Disorders*, vol. 15, no. 4, pp. 287–297, 2001.
[24] A. S. Radomsky, A. J. Ouimet, A. R. Ashbaugh, M. R. Paradis, S. L. Lavoie, and K. P. O'Connor, "Psychometric properties of the french and english versions of the claustrophobia questionnaire (clq)," *Journal of Anxiety Disorders*, vol. 20, no. 6, pp. 818–828, 2006.
[25] R. Shadmehr and S. P. Wise, *The computational neurobiology of reaching and pointing: a foundation for motor learning*. MIT press, 2004.
[26] P. S. Foundation, "Python language reference, version 3.9." Available at http://www.python.org, accessed year of 2021, 2021.
[27] T. A. Stoffregen, B. G. Bardy, L. Smart, and R. Pagulayan, "On the nature and evaluation of fidelity in virtual environments," *Virtual and adaptive environments: Applications, implications, and human performance issues*, pp. 111–128, 2003