Statistics in Motion

A Python based Analysis of the 2 Minute Step Test
Available on GitHub: joerdisstrack/2MST_Statistics_in_Motion

Introduction and Motivation

Since the COVID-19 pandemic, at-home workouts have become increasingly popular, supported by a variety of digital resources like YouTube videos, training plans, and fitness apps. However, performing these exercises effectively without real-time feedback poses a significant challenge for many test takers. For older adults, in particular, maintaining physical independence is crucial, yet they often face difficulties accessing fitness specialists or going outside, making easily accessible feedback essential for staying active and healthy at home.

The 2-Minute Step Test (2MST) in particular is a functional fitness test used to assess aerobic endurance. It involves measuring the number of full steps (knee lifts) completed within two minutes, where a full step is defined as raising each knee to a point midway between the right kneecap (patella) and right hip bone (iliac crest). This test is commonly used for older adults to evaluate their fitness level, particularly lower-body strength and endurance, which are critical for maintaining functional independence. This project aims to bridge this gap by offering personalized feedback on test takers' performance during the 2MST using a python based heuristic.

As a data science project (6 ECTS) within the Master's Program of Social and Economic Data Science at the University of Konstanz, this project is focused on creating an analysis script that evaluates the 2-Minute Step Test using coordinates generated from video files in mp4 format.

Task Description

The program should begin by generating data from smartphone video recordings, transforming this data into coordinate form, and parsing it into a workable CSV format. In the preprocessing phase, it should convert frame-generated timestamps into precise time intervals measured in milliseconds. To ensure the accuracy of the data, the program must apply Gaussian smoothing with a sigma of 2 to reduce the impact of minor positional fluctuations caused by the high sensitivity of the coordinate transformation. This is essential to avoid misleading results during subsequent analysis.

The program should then compute velocity by calculating the differences between timestamps, ensuring a uniform distribution of data points. It must establish a knee threshold and dynamically calculate a ground interval to determine whether a foot is solidly planted while the opposite leg moves, ensuring that only valid steps are counted. The program should define step sequences based on the 2-Minute Step Test (2MST) criteria, incorporating the knee threshold and ground interval with a velocity check to ensure only accurate step sequences are recognized.

Additionally, the program should check for proper form by analyzing cross-body coordination, ensuring that the right leg and left arm, and vice versa, move in synchrony. The data should then be segmented into four equal segments of 30 seconds each to assess the consistency of performance over time, particularly for elderly participants who may experience a decline in performance. Finally, the program should use the performance metrics obtained from this analysis to provide individualized feedback, helping test takers improve their form and endurance.

Theory

The 2-Minute Step Test (2MST) is a simple, yet effective tool used by sports scientists and clinicians to reliably evaluate exercise endurance, particularly in older adults (Rikli & Jones, 1999a), and individuals with mobility impairments (Akkan et al., 2024). The test involves the participant stepping in place for two minutes, raising their knees to a predetermined height, usually at the midpoint between the patella and iliac crest, while counting the number of steps taken. In some settings, instead of raising both knees consecutively, the test takers are invited to step onto a small ladder or stepper instead (Haas et al., 2017). The 2MST was initially developed as part of a broader effort to create a functional fitness test for older adults, as outlined by Rikli and Jones (1999a). Their work aimed to assess physiological parameters supporting physical mobility, such as lower-body strength and aerobic endurance, in order to prevent or delay the onset of physical frailty.

The test's widespread adoption in clinical and rehabilitation settings stems from its practicality, minimal equipment requirements, and its utility in both assessing physical capacity and tracking improvement over time. Recent studies have investigated its validity and reliability in various populations. For example, a study by Bohannon and Crouch (2019) highlighted its utility in older adults as a measure of exercise capacity, particularly its strong correlation with other endurance measures like the 6-Minute Walk Test (6MWT), to which the 2MST is closely linked. Originally developed as an alternative to the 6MST (Rikli & Jones, 1999a), in cases where test takers experienced severe physical impairments like heart failure, the 2MST serves as a reliable replacement that is easier to complete and provides the same level of information on both aerobic exercise capacity and cognitive processes as the 6MST (Michael L. Alosco et al., 2012; Wegrzynowska-Teodorczyk et al., 2016). Similarly, a study by Guedes et al. (2015) confirmed the test's accuracy in diagnosing functional capacity among hypertensive elderly individuals, who are affected by chronic health conditions.

Further exploration of the 2MST has focused on its applicability to individuals with specific health conditions, such as those recovering from stroke or with lower-limb musculoskeletal disorders (Wegrzynowska-Teodorczyk et al., 2016). A 2024 study by Ishigaki et al. (2024) investigated the test's reliability and validity in these populations, finding it to be a valid tool but noting limitations in congruence with other endurance tests like the 6MWT. The study called for further validation with larger sample sizes and highlighted the potential for bias during retesting. Similarly, Nogueira et al. (2021) assessed the reliability of the 2MST in both active and sedentary adults, finding high intrarater and interrater reliability but low accuracy in differentiating between these two groups.

There is an ongoing academic debate surrounding the 2MST, particularly regarding its reliability across different populations with different health impairments and its congruence with other measures of endurance. As research continues, fine-tuning the test's thresholds and validation across diverse cohorts will further solidify its role as a functional and accessible measure of physical fitness that can be easily conducted by test takers themselves from the comforts of their homes.

2MST Algorithm

Data

To test the performance of the algorithm, four different datasets were generated using the open source and python-based programming libraries: MediaPipe (Lugaresi et al., 2019) and OpenCV (Bradski, 2000). The data generation consisted of first recording one's performance of the 2MST using any recording device and saving the video file to mp4 format. These four files were loaded into a program that overlayed the person in the video with 32 relevant markers and proceeded to estimate coordinates of the marker positions frame by frame. After the entire video was processed, the coordinates were standardized into a range of 0 till 1 for generalizability and saved into a dataframe, which additionally contained relevant metadata like an exact time stamp, the frame number and the frame rate. The four datasets are 3606, 3582, 3594 and 3685 frames long and aim to resemble different performances of older test takers, ranging from consistent performances with well-defined and correctly executed step sequences to performances that include short interludes of test takers finding their balance and raising their knees to insufficient heights. Datasets 1 and 3 specifically represent the first category, whereby dataset 1 depicts the test taker performing at a faster pace. Additionally, dataset 3 includes a blurred segment of three seconds, for which coordinates could still be obtained successfully. Datasets 2 and 4 include inconsistent movement, minor breaks and video frames in which the test taker hovers their feet over the ground for diagnostic purposes.

Data Generation

To analyze a test taker's performance of the 2MST based on mp4 files, a full code workflow is provided on GitHub. To overlay the mp4 file with landmarks and obtain coordinates, a local Python environment is preferred due to package dependencies of the MediaPipe library. After downloading the code, the path files must be adapted to the test taker's video and upon execution, the code extracts and processes landmark coordinates frame by frame, while ensuring that the frames per second (FPS) value is valid to compute accurate timestamps for each frame. The utilized Pose Model offers 32 landmarks, covering essential joints and limbs of the human body; a detailed description can be obtained at Google AI for Developers.

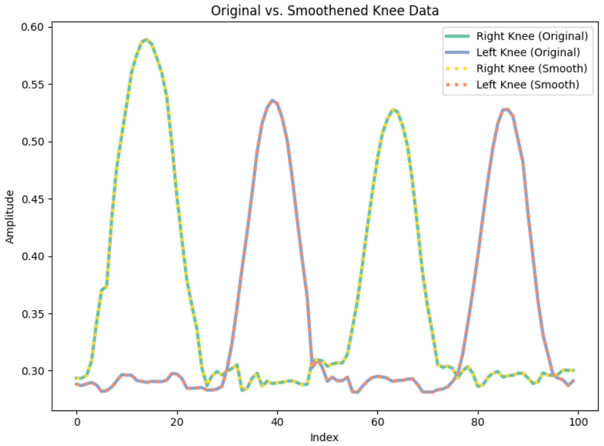

Figure 1: Amplitudes of the right knee over time before and after applying a Gaussian filter

For successful pose estimation, each frame is converted from BGR to RGB format and the pose detection model then processes the frame to identify body landmarks. For each landmark x, y, and z coordinates are extracted and stored in a list and a valid timestamp for each frame is computed based on the frame number and FPS. After processing all frames, the extracted landmarks and timestamps are organized into a Pandas DataFrame. Each landmark's x, y, and z coordinates are labeled with appropriate column names. Finally, to allow for generalization and comparability of results independent from test takers' height, the y-coordinates are standardized to fall within the range [0, 1]. The processed data, including timestamps and standardized values, are saved to a CSV file. Upon completion, a status message confirms the successful extraction, processing, and saving of the data. This workflow is easy to use and allows anyone with access to a local Python distribution to assess their own performance of the 2MST anywhere in the world.

Preprocessing

The generated data requires minimal preprocessing to smooth out potential noise using a Gaussian filter with a sigma value of 2. The purpose of this smoothing is to reduce the noise caused by the sensitivity of the transformation software, which may overreact to minor position shifts during moments where test takers rebalance themselves for better stability, which introduces spikes in the coordinate data. To preserve the integrity of the data without over-smoothing a moderate sigma of 2 is chosen, ensuring that significant motion patterns are retained while eliminating minor fluctuations. The Gaussian smoothing is applied to all x, y, and z coordinates of the pose landmarks. This method leverages the `gaussian_filter1d` function from the SciPy library for efficient one-dimensional filtering (Virtanen et al., 2020). As Figure 1 indicates, minimal smoothing is required when coordinates are generated using MediaPipe and OpenCV as the data contains only minor noise.

Tracking Steps

A valid step is defined as the following sequence of events: First, the right knee must be lifted higher than a predefined threshold while the left foot is solidly placed on the ground, which must then be reversed by lifting the left knee while the right foot is placed solidly on the ground. Ground contact, however, is not clearly defined in most test instructions. After consulting with a sports scientist, this project will consider solid ground contact as at least the toes of any foot to be placed solidly on the ground during the movement of cross diagonal leg, without any jumping or skipping interludes. To successfully track steps, a knee threshold will be computed, moments where either foot is placed on the ground will be detected and to further enhance the ground detection the velocity of relevant landmarks will be computed.

Computing the Knee Threshold

The function `find_knee_threshold`, is designed to calculate a threshold marker based on the vertical (Y) coordinates of the right knee and hip. This threshold mirrors the height threshold that is ascertained by choosing the point between the right hip and knee during real-world assessments of the 2MST. Mathematically, the threshold is obtained by defining the Euclidean distance between the right knee and hip marker at the index value of the global maximum of the right foot marker. At this index position, the right leg is the most extended throughout the entire test performance and it is thus ideal, to evaluate the length of the extended leg. The knee threshold is obtained by averaging the Euclidean distance and differentiates between valid steps where the knees must be lifted above this minimum height and those that are considered non-valid attempts.

Velocity and Ground Contact Detection

Since the detection of ground contact during any performance of the 2MST requires a second proctor while the first oversees the knee movement, the algorithm aims to provide an autonomous functionality that assesses if at least the toes of either foot are placed solidly on the ground by constructing an interval of ground contact. The code detects local maxima and minima in both feets' y-coordinates, calculates ground contact intervals based on the foot and ankle marker's position, and enhances these intervals using the foot's velocity to distinguish between active and inactive phases. The ground contact intervals are dynamically calculated using both initial stances during the first three seconds before the first knee peak and movement data, ensuring accuracy even in cases where there might be an offset between either feet. Such an offset is present in cases, where a test taker bends either leg slightly, which would lead to one foot's y-coordinates to initially be set higher than those of the other foot. Such an offset is also possible in instances, where the test taker's initial stance is relaxed and neither leg is fully extended before attempting the 2MST. The algorithm thus tracks the y-coordinates for both feet individually over the first three seconds, unless an attempt at a valid knee peak is made, in which case the time period of the initial stance is cut short. This initial stance, where both feet must be placed solidly on the ground will form the basis of any ground contact interval.

Next, the `find_all_minima` function locates all local minima in the test takers' feet position during the test. These index values specifically mark further points of assured ground contacts and serve as indicators of the range of values a ground contact interval must span. To achieve this, the data is inverted, allowing the use of the same peak detection method utilized to inspect the knee position over time. The function then applies a threshold based on the mean height of the ankle marker, ensuring that only significant minima (below this threshold) are considered valid. This excludes instances where a test taker might hover their feet above the ground, where the low velocity might otherwise indicate ground contact through a phase of inactivity.

The function `create_ground_contact_interval` then sets up an interval to detect ground contact based on the identified minima of the feet and initial stance data. The lower bound of the interval is fixed at zero -the minimum range of the coordinates- while the upper bound is dynamically determined using the largest minima or the initial stance value, whichever is greater. Finally, the function `detect_ground_contact` identifies moments when the foot is placed on the ground by combining the foot marker's position with its velocity. In a dataset where time intervals between consecutive points are uniform, the instantaneous velocity can be approximated using finite differences. The velocity of either foot can be computed by calculating the difference in position between two consecutive points, divided by the constant time interval between them. The algorithm utilizes the velocity to find phases of inactivity by setting a velocity threshold based

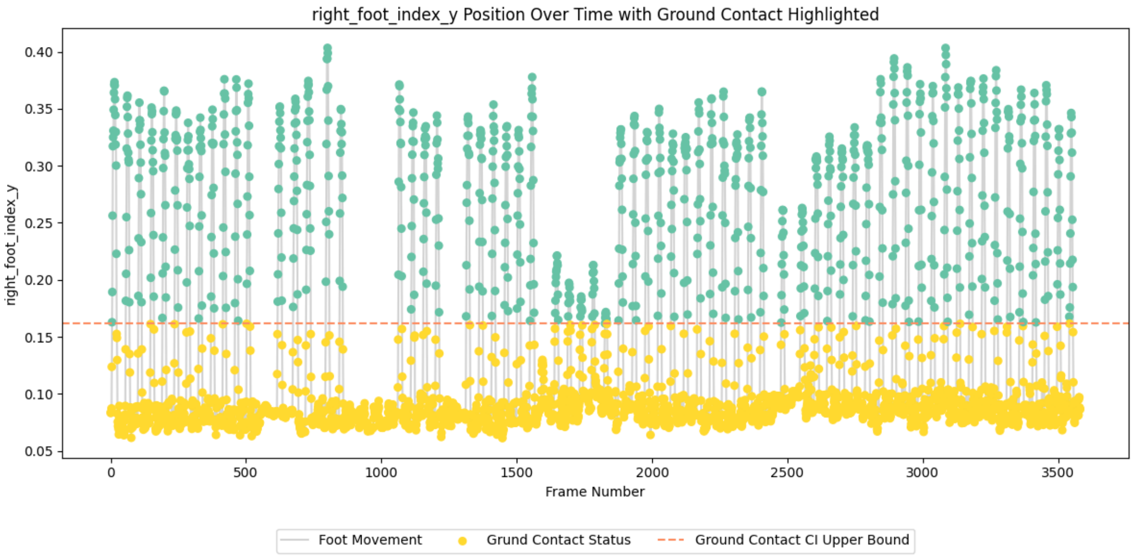

Figure 2: Detection of ground contact for the right foot

on the 95th percentile to consider any data point within the ground contact interval with a velocity of approximately zero as on the ground.

Step Analysis

Once the conditions for valid step sequences can be tracked using the height threshold and the ground contact detection, the algorithm continues to identify valid step sequences. Specifically, the function `validate_step_sequence` identifies valid steps by analyzing right and left knee peaks that surpass the minimum height requirement along with ground contact information. To ensure that the very last step sequence is checked before the algorithm terminates, a flag is raised at the last right knee peak to account for the final segment during analysis, effectively ensuring that the last step is not missed due to indexing functionalities.

The mechanism works by iterating through each right knee peak and searching for the next valid left knee peak that follows it. A step sequence strictly consists of a valid right knee peak that is followed by a left knee peak. Any sequence of consecutively raised right or leg peaks is considered invalid. For each potential step, ground contact for the left foot between the right and left knee peaks is evaluated by assessing an interval of index values centered around the index of the current right knee peak. To complete the assessment, ground contact for the right foot is checked against the index value of the paired left knee peak right knee peaks. If both conditions —ground contact for both feet and height requirement— are met, the step is considered valid and a counter incremented. In the end, the function returns a list of valid step sequences, along with the total count of valid steps detected.

To enhance the step sequence analysis, posture is evaluated by the function `revised_step_sequence_analysis_with_arm_coordination`. This builds on the logic to track steps by adding a check for cross body coordination, which is defined as lifting the opposite arm while the corresponding knee is lifted. For any valid step, this function considers hand movement peaks for both the right and left wrists. Due to the flow of motion during the test, a test taker will typically display a short moment of inactivity during the peak of a knee and its cross diagonally opposed hand. By looping through the identified valid step sequences and searching for local maxima in the y-coordinates of the cross diagonal hand, cross body coordination can be verified: If the right hand is lifted when the left knee is raised, and vice versa, the step is considered to show correct arm-leg coordination. To allow for minor time delays between the peaks, the algorithm searches for hand peaks within a time window of two hundred milliseconds. The function returns the number of coordinated sequences, including both the total step count and the correct coordination count. Together, these functions allow for the detection of valid steps and provide an analysis of good posture during movement.

Consistency Analysis over Segments

In consultation with a sports scientist, the test taker's performance mainly consists of the number of valid steps that were taken during the two minutes of the test, but to provide useful feedback that further discusses performance, a segmentation analysis of the data can be provided. The feedback thus additionally includes the number of breaks that are at least three seconds long, the average heights to which both knees were lifted, variation in these heights for either side individually as well as across both legs, together with the number of valid steps under correct posture that were taken during each segment and the time between knee peaks. Supplying additional comments on the performance can help identify imbalances due to different heights to which the knees are lifted, longer average timespans between peaks of either knee or even breaks that were taken due to imbalance.

To split the data into four equal segments of 30 seconds, the function `find_segment_buffer` is used to calculate a buffer period specific to the data provided by the test taker that based on the average and standard deviation of the time each foot spends in the air, ensuring accurate segment transitions. This is of great importance to prevent overesti-

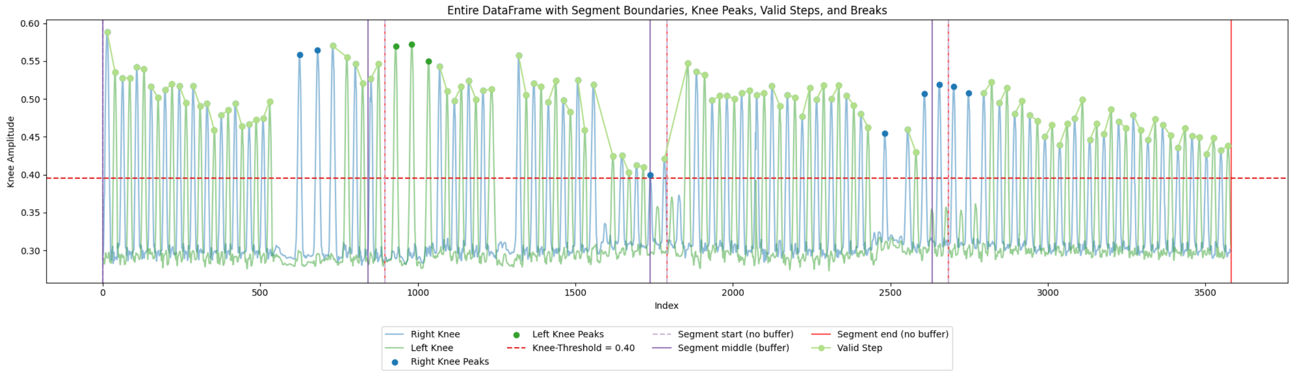

Figure 3: Caption

mating the number of valid steps: If a step is initialized in one but ended in the following segment, defining a meaningful buffer between segments allows to search for an unpaired right knee peak at the very end of one segment in case of the ensuing segment beginning with an unpaired left knee peak. The buffer is computed by considering all attempts at knee peaks, even ones that are invalid by the height requirement as to not distort the average time between phases of activity and inactivity.

Once the data is split, each segment undergoes detailed analysis through the `analyze_segment` function. This function applies the full analysis to each segment by identifying knee peaks and ground contact for both legs to find step sequences and cross body coordination by analyzing hand and arm movements. Additional metrics such as the amplitude of knee peaks and the variation between consecutive peaks are obtained and stored per segment. Figure 3 displays the results of the segment analysis; clearly depicting paired steps over all four segments. After processing all segments, the `compute_consistency_metrics` function aggregates the results, calculating average and standard deviation for the number of steps, cross body coordination, and the variation in knee peak amplitudes. These consistency metrics are then stored in a dictionary and used to form an overall consistency score, which reflects the test taker's performance in terms of step accuracy, coordination, and movement uniformity across all four segments.

Scoring

The 2MST is conventionally scored by the number of valid steps that were completed over the duration of the exercise. To allow for a meaningful interpretation of these results, benchmark data is used to enhance the understanding of the test taker's efforts and the achieved number of steps. In consultation with a sports scientist, a clinical validation study is required in the future before any further attempts at deriving a relevant scoring system that goes beyond the simple determination of the number of steps can be made.

Benchmark Comparison of Performance

The benchmark data were obtained from Rikli and Jones (1999a), which provide normative data on average step counts for a sample of 7,183 participants between the age range of 60 to 94 years —2,135 men and 5,048 women, which makes it a robust reference dataset. Due to limitations in the available age ranges, supplementary age ranges have been interpolated by inspection of linear trends in the mean scores. Table 1 displays both the original benchmark data, as well as the interpolated additional age ranges on gray background. Noticeably, the standard deviations of the mean number of steps are high: This is attributed to the still small sample sizes of the specific demographic subgroups. The performed benchmark comparison thus includes a self-placement of the test taker in their demographic group and a following direct comparison to the appropriate mean.

While the data from Rikli and Jones (1999) covered individuals aged 60-94, there were no direct benchmarks for participants aged 95 and above. To address this, linear interpolation was applied to estimate step counts for older age groups up to 120 years. This interpolation is based on the natural decline in physical performance with age, as observed in the 60-94 age group. By calculating the rate of decline between each consecutive age group, step count thresholds were extended for individuals aged 95-120. These interpolated values ensure that even test-takers of advanced age are fairly evaluated based on reasonable and biologically consistent estimates of step counts.

For younger test-takers, additional benchmarks were derived using an estimated rate of incline, as younger test takers are expected to achieve higher step counts on average. Interpolating further benchmark data for younger test takers is appropriate, given that the 2MST has been shown to be reliable and valid in predicting peak oxygen consumption in healthy adults aged 40 to 59 years (Freene et al., 2021), and has been applied to evaluate physical mobility of test takers as young as 18 years (Akkan et al., 2024). Further data by (Nogueira et al., 2021) were used to assess plausibility of the interpolated values for younger age ranges. This opens up a holistic coverage of normative data that allows test takers of any age to make use of the present analysis algorithm.

For older participants, a key focus of the 2MST is assessing whether individuals are "physically independent." This determination is based on thresholds defined by (Rikli & Jones, 2013), which establish minimum step counts necessary to maintain physical independence. For men and women between the ages of 60 and 94, these thresholds progressively decrease with age, reflecting the natural decline in physical capabilities. The interpolated thresholds for those aged 95 and above continue this trend, providing a complete benchmark dataset across all age groups.

Combining these sources of benchmark data—normative data for older adults from Rikli and Jones (1999b, 2013) with interpolated values for those aged 15 to 59 as well as

Table 1: Interpolated average count and standard deviation of steps by age and gender based on Rikli and Jones (1999a); N = 7,183; age range = 60-94 years, including normative data for men (n = 2,135) and women (n = 5,048)

Age Range	Start Age	Steps (Male)	SD (Male)	Steps (Female)	SD (Female)
15 - 19	15	119	24	110	22
20 - 24	20	117	24	108	22
25 - 29	25	115	23	105	23
30 - 34	30	112	23	103	23
35 - 39	35	110	22	100	23
40 - 44	40	108	22	98	23
45 - 49	45	106	22	96	24
50 - 54	50	103	21	93	24
55 - 59	55	101	21	91	24
60 - 64	60	101	21	91	24
65 - 69	65	101	23	90	26
70 - 74	70	95	23	84	25
75 - 79	75	91	27	84	24
80 - 84	80	87	24	75	23
85 - 89	85	75	24	70	22
90 - 94	90	69	26	58	21
95 - 99	95	64	27	53	21
100 - 104	100	60	27	49	20
105 - 109	105	55	28	44	20
110 - 114	110	51	29	39	19
115 - 119	115	46	30	34	19
120 - 124	120	42	30	30	18

95 and above results in a comprehensive dataset. This unified framework allows the assessment of participants across a broad age range, from 15 to 125 years of age. Most importantly, it ensures that test-takers are compared fairly against their peers, accounting for differences in age, gender, and physical activity levels.

Personal Feedback

To complete the analysis, personal feedback is provided that aims at highlighting both performance achievements and areas for potential improvement. It provides specific insights for each major aspect of the 2MST, ensuring that the participant understands their results in terms of step count, coordination, knee height, and timing consistency. Each section is designed to be informative and encouraging, with performance indicators either confirming consistency or pointing out areas for refinement. The function `feedback_steps` calculates the total and average steps per segment and uses z-scores to assess if the performance is within a normal range. If a segment shows significant deviation (beyond 2 standard deviations), it provides feedback to help the test taker understand how their step count compares to expected norms. This analysis gives the participant insight into their pacing consistency across the test. To enhance insights the function `feedback_coordination` evaluates how well the test taker maintained proper arm-leg coordination during the steps. Z-scores are used to check the regularity of coordination in each segment. If significant deviations occur, feedback alerts the test taker, emphasizing the importance of maintaining synchronized movement. Additionally, the function includes a correlation between steps and coordination to demonstrate how well these metrics align. Segment-specific ratios between steps and coordination are provided. The function `feedback_knee_height` examines the height of the participant's knees during the steps, focusing on both the average knee height and the variation within and across segments. Little variation in knee height suggests consistent performance, while more variation indicates inconsistency and can serve as an indicator for balance impairments, which is of strong importance for older test takers. The feedback points out where knee movements are stable or need improvement, offering specific suggestions when the variation exceeds expectations. Finally, the feedback concludes by interpreting the consistency of movement using the function `feedback_time_between_peaks`: It identifies the time intervals between right and left knee peaks for each segment and further flags any significant variations in timing, offering feedback on the rhythmic consistency of the test taker's steps. Together with a count of the breaks taken over the entire duration of the test, this allows for a detailed performance evaluation and can be discussed together with a healthcare specialist to judge the test taker's balance, stamina and ability to perform the 2MST.

Results and Discussion

Quantitative Performance Evaluation

To evaluate the algorithm, precision, recall, F1 macro and accuracy were computed for each detection task and both sides of the body separately. The algorithm's predictions are compared against human-based labels, which were obtained through self-annotation. The performance metrics for knee peak detection indicate a high level of accuracy across all datasets, with the majority of both left and right knee peaks being correctly detected. For left knee detection, precision is consistently perfect across all datasets with a value of 1.0. Thus, every detected left knee peak was a true

Table 2: Interpolated cutoff threshold values by age and gender based on Rikli and Jones (2013); N = 2,140; age range = 60-94 year

Age Range	Start Age	Threshold (Male)	Threshold (Female)
60 - 64	60	106	97
65 - 69	65	101	93
70 - 74	70	95	89
75 - 79	75	88	84
80 - 84	80	80	78
85 - 89	85	71	70
90 - 94	90	60	60
95 - 99	95	53	55
100 - 104	100	47	49
105 - 109	105	40	44
110 - 114	110	34	39
115 - 119	115	27	34
120 - 124	120	21	28

positive, and no false positives were identified. This level of precision strongly indicates high reliability of the algorithm, which is important given the rare nature of knee peaks as a detectable event compared to every frame, in which the knees are not lifted to a local maximum that lies above the height threshold.

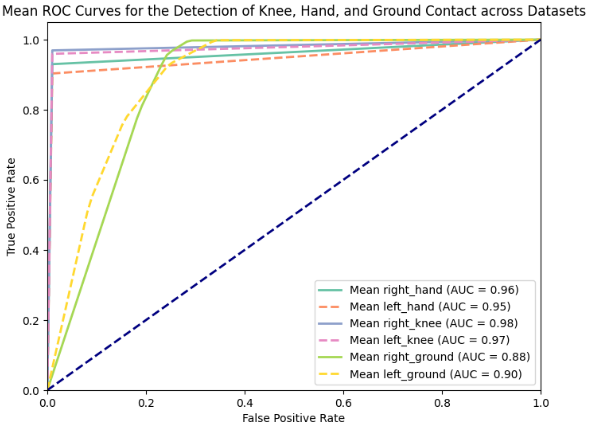

Figure 4: Mean ROC and AUC for knee and hand peak as well as ground detection

However, recall values for left knee detection are slightly lower, ranging from 0.953 to 0.967. This indicates that a small number of true peaks were missed. Despite these missed detections, the algorithm still successfully captured over 95% of the true left knee peaks. Accuracy for left knee detection is very high, with values between 0.9989 and 0.9995, which shows that the vast majority of the data, including both events and non-events, were classified correctly. The F1 scores, which range from 0.976 to 0.983, represent a strong balance between precision and recall. This high level of performance is further reflected in the confusion matrices, which reveal very few false negatives and no false positives.

For right knee detection, precision remains excellent, with perfect results with one exception of 0.976 across the datasets. The minimal number of false positives demonstrates that the algorithm performs similarly well in identifying right knee peaks. Recall for right knee detection varies slightly more than for the left knee, ranging between 0.956 and 0.975. This suggests that between 96% and 98% of the true right knee peaks were detected, despite the left leg partially covering the right leg in datasets 1 to 3, which makes detection more difficult.

As with the left knee, accuracy remains high, with values between 0.9989 and 0.9995, further corroborating the reliability of the algorithm. F1 scores for the right knee are equally strong, ranging from 0.976 to 0.987, indicating the algorithm's robust performance in detecting and classifying right knee peaks correctly. The confusion matrices show slightly more false positives for the right knee than for the left, but these false positives are still at very low levels, with only a few true peaks being missed.

In summary, the algorithm achieves near-perfect precision and high recall. This is especially important because knee peak events are rare, and high precision ensures that false alarms are minimized, preventing distortions in the analysis. The slightly lower recall rates suggest that a small number of true events are missed. This is likely due to personal oversight in listing the actual video frame, at which the local knee maximum occurs during the labeling process. In some cases where several frames looked similar on video, the detection of the exact occurrence of the local maximum where likely noted incorrectly. This is common for human labeling of video materials and given that the detection algorithm relies on SciPy's `find_peaks` functionality (Virtanen et al., 2020), the results suggest perfect detection of knee peaks across the datasets. The high F1 scores further confirm that the algorithm strikes a balance between precision and recall, making it highly effective.

For hand movement detection, the algorithm also shows slightly more variability in recall compared to knee detection. Precision for left hand movement detection is perfect, consistently reaching 1.0 across all datasets, indicating that no false positives were present. Recall, however, ranges from 0.873 to 0.923, meaning that between 87% and 92% of true left hand movements were detected. This slightly lower recall suggests that a higher proportion of true events were missed compared to knee detection. Despite this, accuracy values remain high, ranging from 0.9961 to 0.9980, and F1 scores, which range from 0.932 to 0.960, reflect a strong balance between precision and recall. The confusion matrices show a small number of false negatives, with only a few left hand movements being missed per dataset.

For right hand detection, precision remains nearly perfect,

Table 3: Performance Metrics for Knee Peak Detection

Dataset	**Left Knee**				**Right Knee**			
	Precision	**Accuracy**	**Recall**	**F1**	**Precision**	**Accuracy**	**Recall**	**F1**
dataset_1	1.000	0.999	0.954	0.976	0.976	0.999	0.976	0.976
dataset_2	1.000	0.999	0.953	0.976	1.000	0.999	0.957	0.978
dataset_3	1.000	0.999	0.962	0.981	1.000	0.999	0.974	0.987
dataset_4	1.000	0.999	0.967	0.983	1.000	0.999	0.969	0.984

Table 4: Performance Metrics for Hand Peak Detection

Dataset	**Left Hand**				**Right Hand**			
	Precision	**Accuracy**	**Recall**	**F1**	**Precision**	**Accuracy**	**Recall**	**F1**
dataset_1	1.000	0.997	0.892	0.943	1.000	0.998	0.923	0.960
dataset_2	1.000	0.998	0.923	0.960	0.988	0.997	0.914	0.950
dataset_3	1.000	0.996	0.874	0.933	1.000	0.998	0.932	0.965
dataset_4	1.000	0.998	0.920	0.959	1.000	0.998	0.948	0.973

with values between 0.988 and 1.0. Recall is higher for right hand movements compared to the left, ranging from 0.913 to 0.948, meaning that over 91% of true right hand movements were detected. The accuracy for right hand detection is consistently high, between 0.9975 and 0.9983, confirming the algorithm's reliability. F1 scores for right hand detection are strong, ranging from 0.949 to 0.973, indicating a good balance between precision and recall. The confusion matrices reveal very few false negatives and minimal false positives, with only a small number of missed right hand movements.

The slightly lower values for precision and recall are likely caused by small pauses in hand movement at the local maxima over several consecutive video frames. This makes the detection of the true local maximum very difficult to perform during the human labeling process.

Finally, in the detection of foot ground contact, the algorithm demonstrates strong overall performance, though with greater variability compared to the knee and hand detection tasks, particularly in precision. For left foot ground contact detection, precision ranges from 0.858 to 0.959, meaning that between 85% and 96% of predicted left foot ground contacts were correctly identified as true positives. Recall, however, remains high, ranging from 0.994 to 1.0, showing that almost all true left foot ground contacts were detected. The accuracy for left foot detection ranges from 0.8887 to 0.9684, and F1 scores range from 0.9236 to 0.9769. Confusion matrices reveal more false positives than false negatives, particularly in datasets two and four. Here, the number of incorrect predictions is more noticeable due to failed attempts at valid steps and video frames, where the test taker hovered their foot over the ground rather than moving it consistently up and down.

For right foot ground detection, precision ranges from 0.878 to 0.918, meaning that between 87% and 92% of predicted right foot ground contacts were correct. Recall remains high, ranging from 0.996 to 1.0, and accuracy values range from 0.9047 to 0.9374. F1 scores are solid, ranging from 0.934 to 0.956, despite the presence of some false positives. The confusion matrices reveal that false positives are again more common than false negatives, but overall, the algorithm performs well in capturing true foot ground contacts.

Figure **??** illustrates that for datasets 1, 2 and 3 the AUC of the ground contact detection of the left foot is consistently higher than that of the right foot: This is because the right leg was partially covered by the left leg as all videos were taken from a direct profile rather than an angled perspective. All in all, the ground contact detection can be fine-tuned further, to reduce the number of false positives and be more precise about overall ground contact detection. It is important to note, however, that the algorithm's functionality for step detection is not affected by the slight overestimation of frames in which feet are placed solidly on the ground: The function that tracks valid step sequences searches for an interval of consistent ground contact of the opposite foot around valid knee peaks, and thus does not require overestimation to be non-present for the correct detection of knee peaks and step sequences. Lastly, the present overestimation was caused by overly sensitive data during early development: An initial dataset used in the first stage of development showed local maxima of strong amplitude in the y-coordinates of both feet during moments of consistent ground contact. These local maxima crossed the height of the test takers' ankles at times and were caused by an overly sensitive coordinate generation that captured moments of the test taker readjusting their balance during the performance. During early development it was therefore necessary to adjust the code to over-sensitivity as otherwise frames with present ground contact would have been missed. However, this is a problem that does no longer persist with the current coordinate generation and thus leaves room for future fine-tuning of the ground detection functionality. Figure 5 depicts the individual ROC and AUC for each of the evaluation datasets.

Qualitative Assessment

An interview with a sports scientist supports the first assessment of the algorithm's performance: From their perspective, an algorithm that analyses performance of the 2MST must correctly identify valid knee peaks and can be further enhanced by detecting moments of ground contact for both feet and posture as expressed by the correct coordination of lifting the right knee and the left arm and hand and vice versa, resulting in local maxima of both knee and opposing hand at the same points in time. The correct count of valid step sequences is the most important factor that should influence test takers' final scores. An analysis of posture and valid ground contact over four different segments of

Table 5: Performance Metrics for Ground Contact Detection

Dataset	Left Ground				Right Ground			
	Precision	Accuracy	Recall	F1	Precision	Accuracy	Recall	F1
dataset_1	0.867	0.905	0.999	0.928	0.878	0.911	1.000	0.935
dataset_2	0.959	0.968	0.995	0.977	0.899	0.921	0.996	0.945
dataset_3	0.924	0.945	0.997	0.959	0.918	0.937	0.996	0.956
dataset_4	0.858	0.889	1.000	0.924	0.878	0.905	0.997	0.934

the performance are interesting but secondary aspects that can enhance the algorithm and the feedback that test takers receive.

At present, the algorithm detects valid step sequences correctly and thus satisfies the task set by sports science. The feedback provides useful further information about performances and can be assessed to improve over time. Future work and clinical validation are required to assess the algorithm's full application in a real-world setting and to potentially derive fine-grained scoring systems. Given the international use of the 2MST, which is based entirely on the count of valid steps taken as the only feature relevant for scoring, the derivation of new scoring systems that utilize and weight different performance features are without comparison and their performance and sensibility would remain without meaningful test.

At the current state, the algorithm fully fulfills the need for a highly precise and automatic scoring system as expressed by a sports scientist.

Limitations and Future Improvements

Yet, the algorithm at hand represents a first version of a program to analyze the 2MST and can be improved further: To begin, the ground contact detection can be improved by refining the algorithm to better distinguish between moments of true contact and minor adjustments in balance. This will help reduce the likelihood of overestimating ground contact, especially as the generation of coordinates has changed and initial oversensitivity in the coordinates of either foot is no longer present.

To accurately describe the algorithm's performance, the labeling process must be revised, by either increasing the number of individuals involved in manual labeling to check inter-coder reliability or opting for an automatic labeling procedure that does not depend on SciPy's `find_peaks` function, as this method is already used by the current algorithm. A more robust labeling system will improve the accuracy and reliability of the data. Similarly, a clinical validation of the detection algorithm must be conducted, to ensure that provided feedback is both accurate and meaningful in real-world scenarios. Feedback from clinical testing will be essential to fine-tune the algorithms for practical applications and confirm their utility in various settings. Finally, the thresholds used in the final report for detecting significant variations can be adjusted, such as knee height variation or peak time standard deviation. Consider making these thresholds customizable, allowing participants to adjust them according to their experience level or specific physical conditions. This flexibility could lead to more personalized and relevant feedback for each individual.

Conclusion

In conclusion, the algorithm excels in detecting valid local maxima in hand and knee peaks across different movement types, with particularly strong precision and high recall rates. For knee and hand detection, precision is near-perfect, and recall is consistently strong, though there is some room for improvement in capturing all true events. Performance is strong even if arms and legs are partially covered or the video quality is low as ROC and AUC verify for each dataset. For foot ground contact detection, while precision is slightly lower due to more false positives, the recall remains very high, ensuring that the algorithm captures nearly all true events. To further refine the algorithm, an improvement in consistency and reliability of the human-annotated labels is required, as imperfect values in precision and recall of some datasets can be caused by incorrectly labeled video frames. Overall, the algorithm's performance is highly effective and well-suited for detecting rare movements with minimal false positives and a high rate of true positive detection. The algorithm at hand allows for automatic and innovative scoring of the 2 Minute Step Test and is among the first programs based entirely on open source programming languages, which makes it easy to use and accessible not only for interdisciplinary research but also test takers at home.

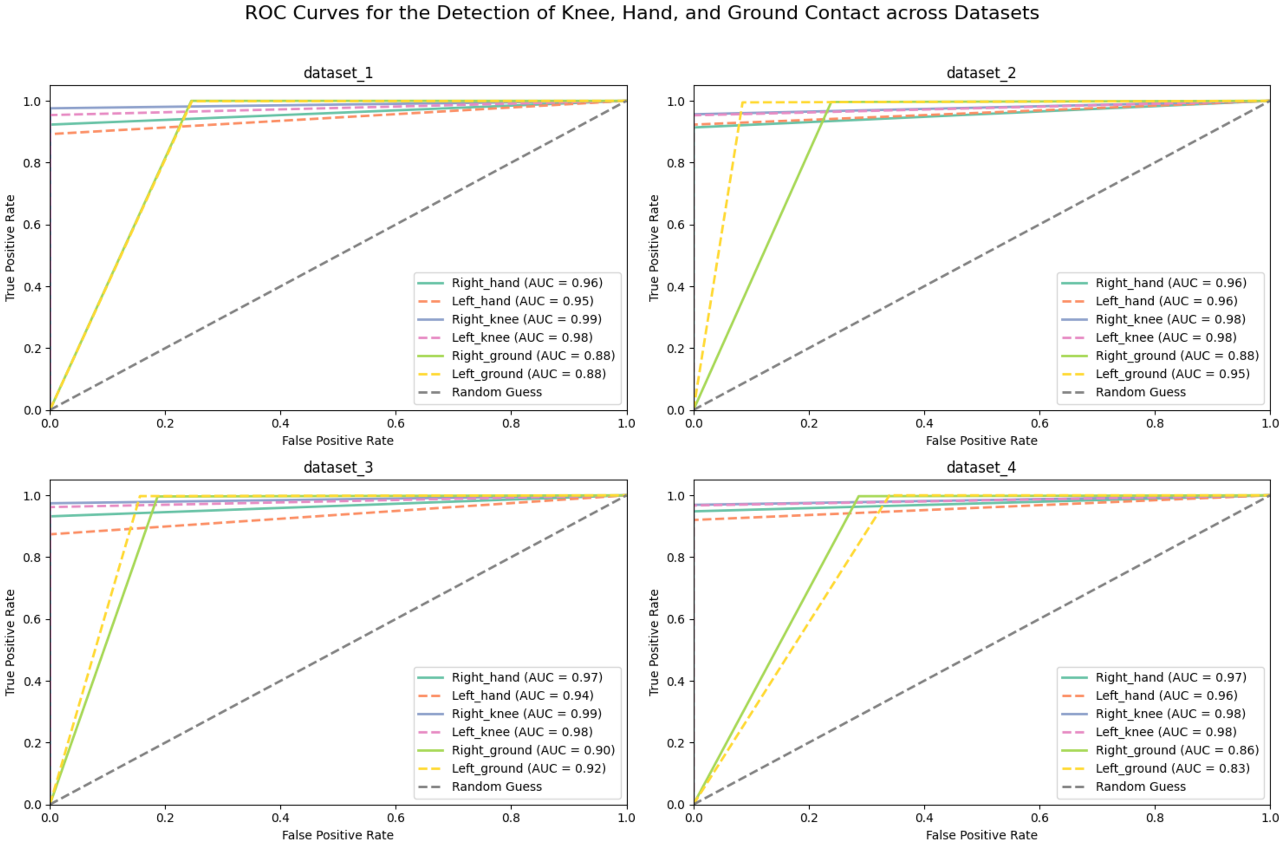

Figure 5: Individual ROC and AUC for each of the four datasets

References

Akkan, H., Kaya Mutlu, E., & Kuyubasi, S. N. (2024). Reliability and validity of the two-minute step test in patients with total knee arthroplasty. *Disability and Rehabilitation*, *46*(14), 3128–3132. https://doi.org/10.1080/09638288.2023.2239141

Bohannon, R. W., & Crouch, R. H. (2019). Two-Minute Step Test of Exercise Capacity: Systematic Review of Procedures, Performance, and Clinimetric Properties. *Journal of Geriatric Physical Therapy*, *42*(2), 105–112. https://doi.org/10.1519/JPT.0000000000000164

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*. https://opencv.org/

Freene, N., Pike, A., Smith, D., Pradhananga, A., & Toohey, K. (2021). Criterion Validity of the Older-adults 2-minute Step Test in Community-dwelling Middle-aged Adults. *Measurement in Physical Education and Exercise Science*, *25*(4), 335–343. https://doi.org/10.1080/1091367X.2021.1904934

Guedes, M. B. O. G., Lopes, J. M., Andrade, A. D. S., Guedes, T. S. R., Ribeiro, J. M., & Cortez, L. C. D. A. (2015). Validation of the two minute step test for diagnosis of the functional capacity of hypertensive elderly persons. *Revista Brasileira de Geriatria e Gerontologia*, *18*(4), 921–926. https://doi.org/10.1590/1809-9823.2015.14163

Haas, F., Sweeney, G., Pierre, A., Plusch, T., & Whiteson, J. (2017). Validation of a 2 Minute Step Test for Assessing Functional Improvement. *Open Journal of Therapy and Rehabilitation*, *05*(02), 71–81. https://doi.org/10.4236/ojtr.2017.52007

Ishigaki, T., Kubo, H., Yoshida, K., Shimizu, N., & Ogawa, T. (2024). Validity and reliability of the 2-min step test in individuals with stroke and lower-limb musculoskeletal disorders. *Frontiers in Rehabilitation Sciences*, *5*, 1384369. https://doi.org/10.3389/fresc.2024.1384369

Lugaresi, C., Tang, J., Doshi, A., Singh, M., Zhang, F., Chang, C.-L., Feng, Q., Lehrmann, T., Mustafa, B., Guo, F., Rong, M. T., Ranjan, A., Chowdhury, R., Andriluka, M., Luo, C., Li, Z., Huang, J., & Gorban, Y. (2019). MediaPipe: A Framework for Building Perception Pipelines. https://google.github.io/mediapipe/

Michael L. Alosco, Mary Beth Spitznagel, Naftali Raz, Ronald Cohen, Lawrence H. Sweet, Lisa H. Colbert, Richard Josephson, Donna Waechter, Joel Hughes, Jim Rosneck, & John Gunstad. (2012). The 2-minute step test is independently associated with cognitive function in older adults with heart failure. *Aging Clinical and Experimental Research*, *24*(5). https://doi.org/10.3275/8186

Nogueira, M. A., Almeida, T. D. N., Andrade, G. S., Ribeiro, A. S., Rêgo, A. S., Dias, R. D. S., Ferreira, P. R., Penha, L. R. L. N., Pires, F. D. O., Dibai-Filho, A. V., & Bassi-Dibai, D. (2021). Reliability and Accuracy of 2-Minute Step Test in Active and Sedentary Lean Adults. *Journal of Manipulative and Physiological Therapeutics*, *44*(2), 120–127. https://doi.org/10.1016/j.jmpt.2020.07.013

Rikli, R. E., & Jones, C. J. (1999a). Development and Validation of a Functional Fitness Test for Community-Residing Older Adults. *Journal of Aging and Physical Activity*, *7*(2), 129–161. https://doi.org/10.1123/japa.7.2.129

Rikli, R. E., & Jones, C. J. (1999b). Functional Fitness Normative Scores for Community-Residing Older Adults, Ages 60-94. *Journal of Aging and Physical Activity*, *7*(2), 162–181. https://doi.org/10.1123/japa.7.2.162

Rikli, R. E., & Jones, C. J. (2013). Development and validation of criterion-referenced clinically relevant fitness standards for maintaining physical independence in later years [Publisher: Oxford University Press]. *The gerontologist*, *53*(2), 255–267.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wegrzynowska-Teodorczyk, K., Mozdzanowska, D., Josiak, K., Siennicka, A., Nowakowska, K., Banasiak, W., Jankowska, E. A., Ponikowski, P., & Woźniewski, M. (2016). Could the two-minute step test be an alternative to the six-minute walk test for patients with systolic heart failure? *European Journal of Preventive Cardiology*, *23*(12), 1307–1313. https://doi.org/10.1177/2047487315625235