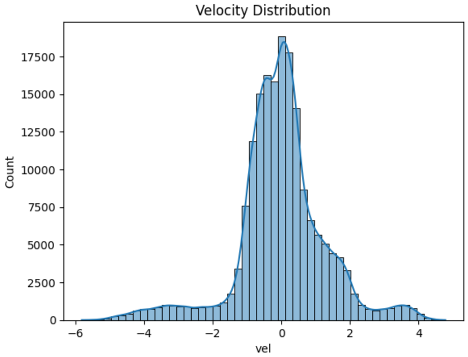
Macro-Level Analysis

1 Feature Engineering

To understand smoothness and variability we engineer derivative features: angular velocity and acceleration. These derivatives provide insights into the rate of change in joint angles over time.

1.1 Velocity

$$\text{velocity} = \frac{\Delta\text{angle}}{\Delta\text{time}} = \frac{\text{angle}_i - \text{angle}_{i-1}}{\text{time}_i - \text{time}_{i-1}} \quad \text{for each (subject, condition, replication, leg, joint)}$$

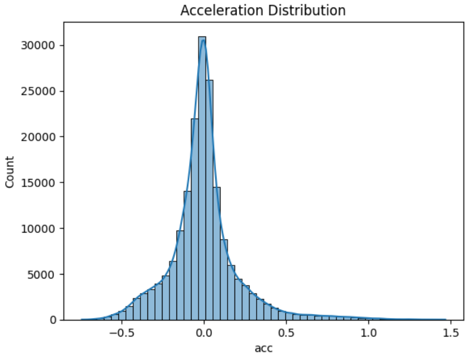


Observations:

- The velocity distribution is approximately normal, centered around zero.
- Slight skewness and extended tails suggest variability in movement intensity.

1.2 Acceleration

$$\text{acceleration} = \frac{\Delta\text{velocity}}{\Delta\text{time}} = \frac{\text{velocity}_i - \text{velocity}_{i-1}}{\text{time}_i - \text{time}_{i-1}} \quad \text{for each (subject, condition, replication, leg, joint)}$$



Observation

- The acceleration distribution is heavily centered around zero, indicating predominantly stable joint velocities in the gait data.

- A right-skewed tail extends to positive values up to 1.5, suggesting infrequent but notable accelerations compared to decelerations.
- Over 80% of observations lie within $[-0.2, 0.2]$, underscoring low variability in angular acceleration across most gait sequences.

1.3 Range (Mobility Span)

$$\text{Angle Range} = \text{angle}_{\text{max}} - \text{angle}_{\text{min}}$$
$$\text{Velocity Range} = \text{velocity}_{\text{max}} - \text{velocity}_{\text{min}}$$
$$\text{Acceleration Range} = \text{acceleration}_{\text{max}} - \text{acceleration}_{\text{min}}$$

1.4 Coefficient of Variation

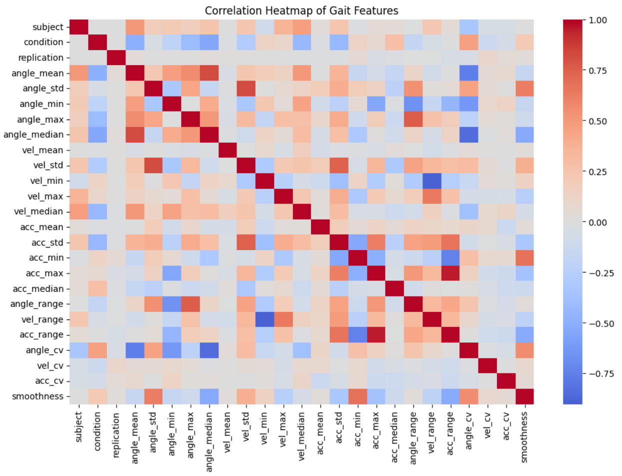
$$\text{Angle CV} = \frac{\text{angle}_{\text{std}}}{|\text{angle}_{\text{mean}}| + \varepsilon}$$
$$\text{Velocity CV} = \frac{\text{velocity}_{\text{std}}}{|\text{velocity}_{\text{mean}}| + \varepsilon}$$
$$\text{Acceleration CV} = \frac{\text{acceleration}_{\text{std}}}{|\text{acceleration}_{\text{mean}}| + \varepsilon}$$

1.5 Smoothness Metric

$$\text{Smoothness} = \frac{\text{velocity}_{\text{std}}}{\text{acceleration}_{\text{std}} + \varepsilon}$$

2 Exploratory Data Analysis (EDA)

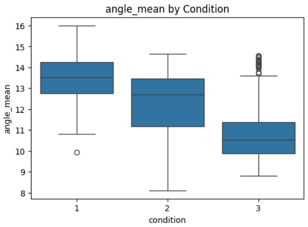
2.1 Feature Correlations



Observations:

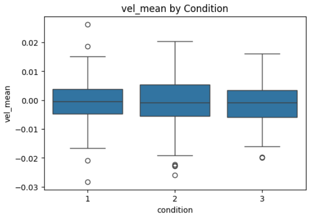
- There is strong positive correlations among angle summary statistics like mean, median, std, and max, Pointing out extra or unnecessary movements.
- Velocity and acceleration features show moderate to low correlations ($|r| < 0.5$) with angle metrics

2.2 Distribution of features across different conditions using boxplots.



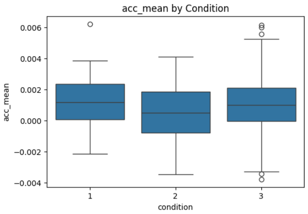
Observation:

- The median angle_mean decreases across conditions, being highest in Condition 1 (approx. 13.5) and lowest in Condition 3 (approx. 10.5).
- Condition 3 has the largest number of high-value outliers, ranging from approximately 14.0 to 14.6.
- Condition 1 has the largest interquartile range (IQR), while Condition 3 has the narrowest, suggesting less variation in its middle 50% of data.



Observation:

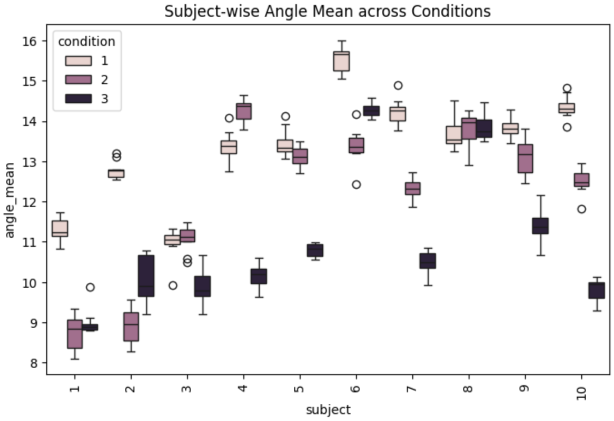
- The median vel_mean is close to zero for all three conditions, implyes that there is not much drift in velocity.
- Condition 2 has the tightest interquartile range (IQR), suggesting that the central 50% of its data is highly concentrated near the zero median.



Observation:

- All three conditions show a positive median acc_mean
- Condition 3 has the largest spread, having the lowest lower whisker (approx. -0.0035) and multiple high-value positive outliers (up to ≈ 0.006).
- Condition 1's distribution is the most symmetric around its median, while Condition 3 is slightly skewed toward positive values.

2.3 Subject-wise Angle Mean across Conditions

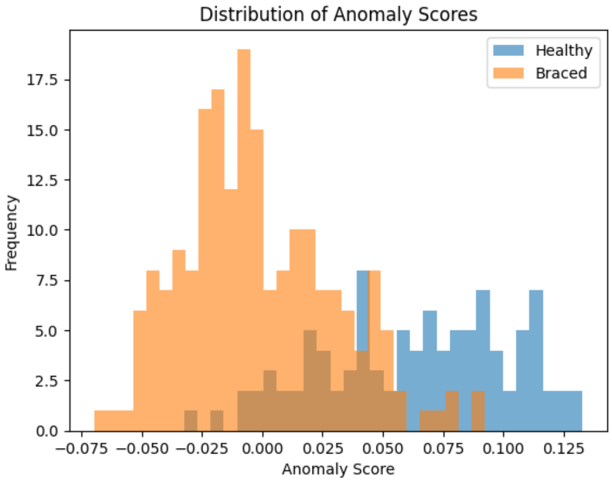


Observation:

- **Condition 3**(ankle braced) consistently shows the **lowest** angle_mean in almost all subjects, usually clustering between 9.5 and 11.0.
- **Condition 1** (unbraced) mostly has the **highest** angle_mean for several subjects.

3 Evaluating Anomaly Score

1. **Model Training and Feature Scaling:** The StandardScaler was fit only on the healthy data (X_healthy) and then used to transform both the training (X_scaled) and test (X_test_scaled) sets. The IsolationForest was trained on X_scaled with $T = 200$ trees (n_estimators = 200) to learn the profile of normal behavior.
2. **Path Length Determination:** For every data point x, the model measures the number of edges it takes to isolate x in each of the $T = 200$ trees. The key metric is the average path length, denoted $E[h(x)]$.
$$E[h(x)] = \frac{1}{T}\sum_{i=1}^{T} h_i(x) \quad \text{where } T = 200$$
Anomalies are isolated quickly (short $E[h(x)]$), while normal points require many splits (long $E[h(x)]$).
3. **Decision Function Score Output:** The scores returned by iso.decision_function(X) are proportional to the negative average path length, $\propto -E[h(x)]$, and are not the normalized score $s(x)$.
$$\text{Score} \propto -E[h(x)]$$
Therefore, a **higher score** (closer to zero or positive) means a **longer** path length, indicating a **normal** point, which aligns with the distribution of healthy_scores.
4. **Prediction Classification:** The contamination = 0.05 parameter implicitly sets the threshold for classification. The prediction assigns $\mathbf{-1}$ (anomaly) to the 5% of points with the lowest scores and **1** (normal) to all points with scores above that threshold.



Observation:

- The **Healthy** data exhibits a higher anomaly score distribution, centered around positive values, consistent with being the training data for normal behavior.
- The **Braced** data shows a clear shift toward lower (more negative) anomaly scores, indicating that these points are more frequently isolated as potential anomalies.
- The two score distributions are **partially separated** with the optimal threshold for distinguishing anomalies likely falling within the overlapping range of 0.000 to 0.070.

4 Evaluating Gait Score

Isolation Forest

The Gait Health Score converts the raw IsolationForest anomaly scores into an intuitive scale from 0 to 100.

1. **Obtain Raw Anomaly Scores (score):** The raw anomaly scores (test_scores) are computed for the braced data using the trained Isolation Forest model (iso.decision_function).

$$\text{score} = \text{iso.decision\_function}(\text{X\_test\_scaled})$$

2. **Normalize Score to $[0, 100]$:** The scores are normalized such that the maximum score (most normal point) maps to 100 and the minimum score (most anomalous point) maps to 0. This normalized result is the Gait Health Score.

$$\text{Gait Health Score(score)} = 100 \times \frac{\text{score} - \min(\text{score})}{\max(\text{score}) - \min(\text{score})}$$

3. **Compute Average Health and Anomaly Count:** The average score for the braced data is calculated (result: 42.15), and the number of samples predicted as anomalies ($-\mathbf{1}$) is counted (result: 118/200).

$$\text{Average Health Score} = \text{mean(health\_scores)}$$

One-Class SVM Comparison

To compare the performance of different anomaly detection techniques, a OneClassSVM model is initiated and its scores are also normalized.

1. **Train One-Class SVM Model:** The OneClassSVM model is trained exclusively on the healthy data (X_scaled) using a radial basis function (rbf) kernel and a nu parameter of 0.05.

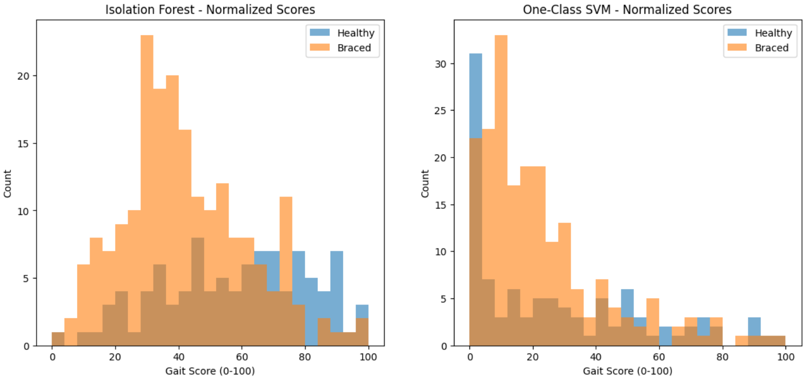
$$\text{Model} = \text{OneClassSVM}(\text{kernel='rbf'}, \nu = 0.05)$$

2. **Obtain SVM Decision Scores:** The decision function scores are calculated for both the healthy and braced data using the SVM model.

$$\text{SVM Scores} = \text{svm.decision\_function}(\text{X\_scaled or X\_test\_scaled})$$

3. **Normalize All Scores for Comparison:** The scores from both models (IsolationForest and OneClassSVM) are normalized to the $0 - 100$ range to allow for a direct quantitative comparison of their anomaly detection output.

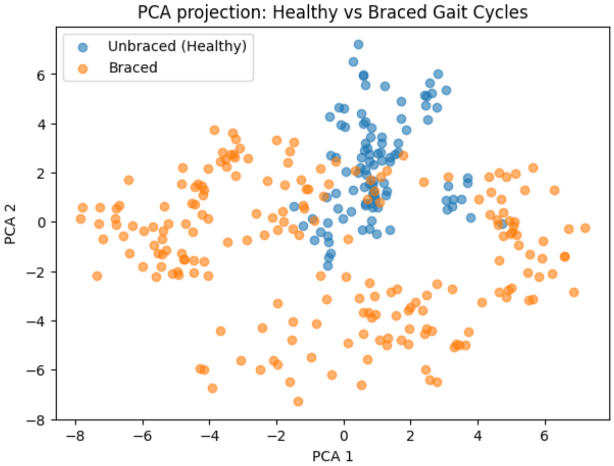
$$\text{Normalized Score} = 100 \times \frac{\text{raw\_score} - \min(\text{raw\_score})}{\max(\text{raw\_score}) - \min(\text{raw\_score})}$$



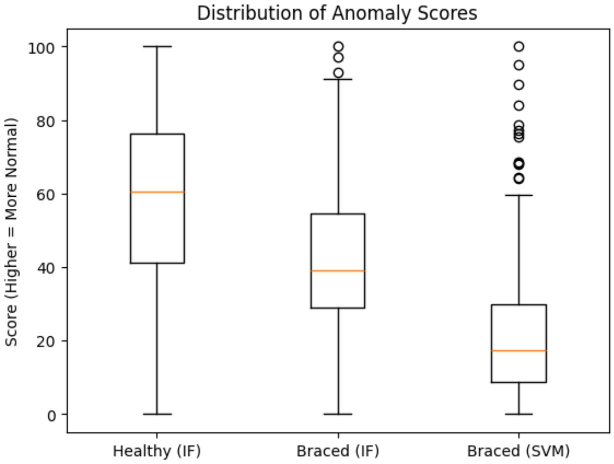
Observation:

- The IsolationForest scores show the Braced data spread relatively widely (from 0 to ≈ 90), with the highest frequency between 20 and 40.
- The OneClassSVM indicates majority of braced data has gait score between 0 to 20
- When compared to the OCSVM, the IF model provides a more nuanced Gait Score distribution for the Braced data, showing a larger proportion of samples scoring above 40.

5 Principal Component Analysis



This overlap explains why anomaly detection struggles—many braced cycles appear “normal,” and some healthy cycles fall near the boundary. Supervised models perform better because they use the full feature space, not just the top two principal components.



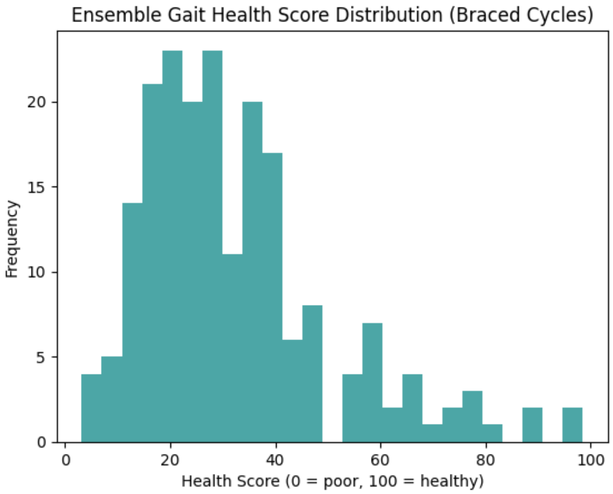
Isolation Forest (IF) shows clear separation between Healthy and Braced

SVM on Braced data appears to have slightly different distribution than Braced (IF)

6 ENSEMBLING GAIT HEALTH INDEX (IF + SVM)

A final, more robust ensemble_score is created by taking the unweighted average of the two normalized model scores (health_IF and health_SVM).

$$\text{ensemble\_score} = \frac{\text{health\_IF} + \text{health\_SVM}}{2}$$



Observation:

- The Ensemble Score distribution is **highly skewed** towards the lower, unhealthy end, with the main concentration of scores falling between 15 and 40.
- The highest frequency peak occurs sharply around the 20 to 30 Health Score range, indicating that the majority of Braced cycles are classified as having poor gait health.

7 Classification (Predicting Gait Condition)

1. **Target Variable Encoding and Data Split:** The original condition (1, 2, 3) is converted to a binary target condition_binary (0 = Unbraced, 1 = Braced). The data is split into training and testing sets based on unique **subject** IDs, ensuring model generalization.

$$y = \begin{cases} 0 & \text{if condition} = 1 \\ 1 & \text{if condition} \in \{2, 3\} \end{cases}$$

2. **Feature Scaling (for Logistic Regression):** The training features (X_train) are scaled using StandardScaler and the same transformation is applied to the test features (X_test) to ensure model convergence and prevent features with larger magnitudes from dominating the linear model.

$$\text{X\_scaled} = \frac{X - \mu}{\sigma}$$

3. **Model Training (Logistic Regression):** A Logistic Regression model is trained on the scaled training data, finding the linear relationship between the features and the log-odds of the gait condition.

$$\log(\frac{P(Y = 1)}{1 - P(Y = 1)}) = \beta_0 + \sum_{i=1}^{22} \beta_i X_i$$

4. **Model Training (Random Forest):** A non-linear Random Forest classifier is trained with n_estimators = 200 on the unscaled training data, building an ensemble of decision trees to capture complex feature interactions.

5. **Model Evaluation and Feature Importance:** Both models are evaluated on the held-out test set (X_test), reporting Accuracy, Precision, Recall, and F1-score. Feature importance is derived from Logistic Regression coefficients (importance = $|\beta|$) and Random Forest impurity reduction.

Performance Summary

- The Logistic Regression model achieved a very high Accuracy of **0.956**, outperforming the Random Forest model (Accuracy : 0.889).
- Logistic Regression demonstrated balanced performance with Precision and Recall of **0.93** for class 0 (Unbraced) and **0.97** for class 1 (Braced).

Micro-Level Analysis

1 Feature Engineering

1.1 Jerk

Jerk shows how quickly acceleration changes, helping us spot sudden or uneven joint movements.

$$\text{jerk}_i = \frac{\text{acc}_i - \text{acc}_{i-1}}{\text{time}_i - \text{time}_{i-1}} \quad \text{for each (subject, condition, replication, leg, joint)}$$

1.2 Symmetry Index (per timepoint)

$$\text{Symmetry Index} = 1 - \frac{|\text{angle}_L - \text{angle}_R|}{|\text{angle}_L| + |\text{angle}_R| + \varepsilon} \quad \text{clipped to } [0, 1]$$

- Measures how similar left and right joint angles are at each timepoint.
- A value close to 1 indicates high symmetry; closer to 0 suggests imbalance.

1.3 Stability Score (per leg)

$$\text{Stability Score} = 1 - \frac{\text{std}_{\text{angle}}}{\text{range}_{\text{angle}} + \varepsilon} \quad \text{clipped to } [0, 1]$$

- Captures how consistent joint movement is over time.
- Lower variability relative to range implies better stability.

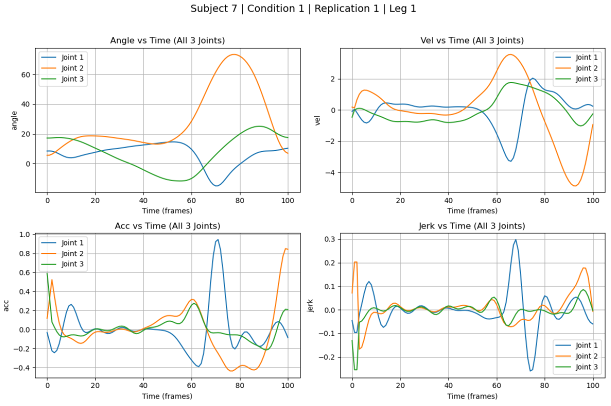
1.4 Gait Score (per leg)

$$\text{Gait Score} = 50 \cdot \text{Symmetry Index} + 50 \cdot \text{Stability Score}$$

- Combines symmetry and stability equally to assess overall gait quality.
- Higher scores reflect smoother, more balanced movement.

2 Exploratory Data Analysis (EDA)

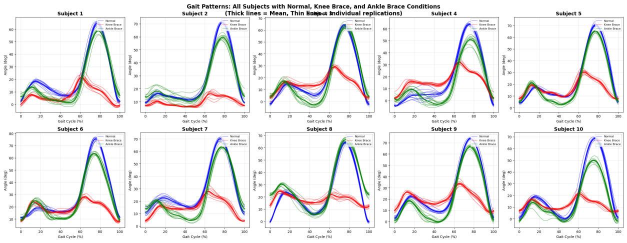
2.1 Joint Dynamics Over Time



Observations (Subject 7, Condition 1, Replication 1, Leg 1) :

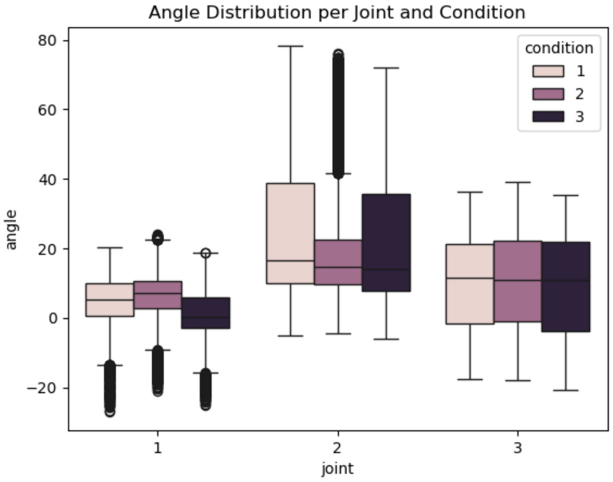
- The **Knee** (Joint 2) shows the greatest range of motion, with its angle changing the most and its velocity having the largest swings.
- The **Ankle** (Joint 1) is the least smooth joint, showing the highest and sharpest peaks in acceleration and jerk, which suggests sudden, rapid changes in muscle activation.
- The **Hip** (Joint 3) has the most stable movement patterns across all four metrics, indicating its role in providing steady, large-scale support during the cycle.

2.2 Gait Patter Across Subjects for Knee joint



- **Normal condition** shows smooth knee movement with clear flexion-extension cycles across most subjects.
- **Knee brace** slightly reduces knee range and smoothness, especially in Subjects 3, 6, and 9.
- **Ankle brace** affects knee motion less directly, but some subjects (like 2 and 5) show altered timing or reduced peaks.

2.3 Angle Distribution per Joint and Condition



- **Ankle (Joint 1)** shows the widest angle spread, especially in the normal condition, with several outliers.
- **Knee (Joint 2)** has moderate variation; knee brace condition slightly lowers the median angle.
- **Hip (Joint 3)** shows the most compact distribution, with ankle brace condition having the tightest range.

3 Anomaly Detection

1. **Feature Selection and Dataset Subsets:** The data is partitioned into Healthy (condition = 1) and Anomalous (condition $\in \{2, 3\}$) subsets.Below are selected as features **X**.

$$\mathbf{X} = \{\text{angle}, \text{vel}, \text{acc}, \text{jerk}, \text{symmetry\_index}, \text{stability\_score}, \text{gait\_score}\}$$

2. **Standardization (Scaling) based on Healthy Data:** A StandardScaler is fit **only** on the Healthy data (X_healthy) to calculate the mean (μ) and standard deviation (σ) of the normal distribution. All subsets are then transformed using these parameters.

$$\text{X\_scaled} = \frac{X - \mu_{\text{healthy}}}{\sigma_{\text{healthy}}}$$

3. **Isolation Forest Training:** An IsolationForest is trained exclusively on the X_scaled data, using n_estimators = 200 and setting the expected anomaly rate (contamination) to 0.05.

$$\text{Model} = \text{IsolationForest}(\text{n\_estimators} = 200, \text{contamination} = 0.05)$$

4. **Anomaly Score and Prediction:** The trained model computes the raw anomaly score for the Healthy (healthy_scores) and Anomalous (test_scores) subsets, and makes binary predictions ($\mathbf{1}$ = Normal, $-\mathbf{1}$ = Anomaly) based on the contamination threshold.

$$\text{Scores} \propto -\frac{\text{Average Path Length}}{c(n)}$$

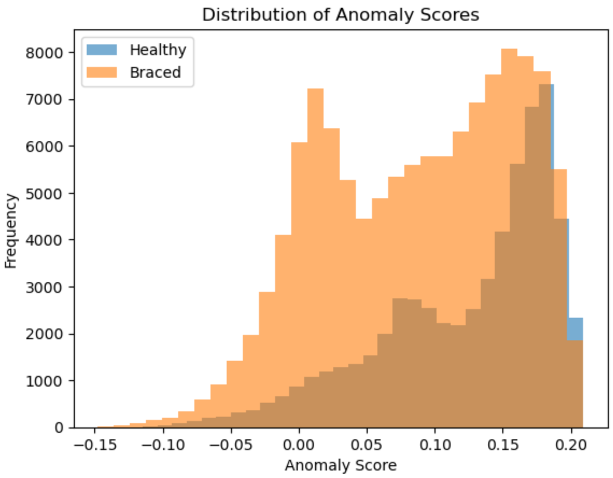


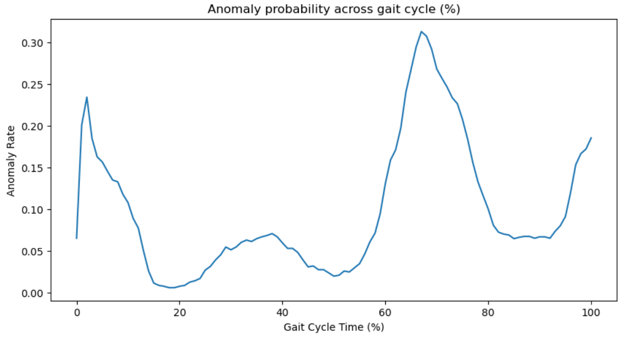
Figure 1: Anomaly score for Isolation forest

- The Healthy data shows a high concentration of scores on the positive side, meaning the model sees them as the most normal.

- The Braced data is widely spread out, with its main peak being closer to zero, indicating it is easier to isolate as anomalous.

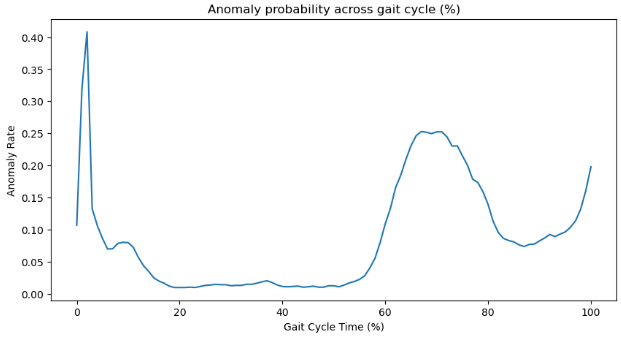
- The two distributions have large area of overlap, making a simple threshold ineffective for separating all healthy and braced samples.

Anomaly probability across gait cycle (%) for IF



- The gait cycle shows two major periods of high anomaly rate: one at the very start (≈ 0-5%) and a much larger peak around the **65%** mark.
- The highest risk of anomaly occurs sharply at about **65%** of the cycle, reaching a peak anomaly rate of just over 0.30 (or 30%).
- The period between 15% and 50% of the cycle is the most stable and least anomalous.

Anomaly probability across gait cycle (%) for OCSVM



- The biggest anomaly risk is at the **0%** mark (start of step), with the rate spiking above **40%**.
- The second, lower risk peak occurs around **65%** of the cycle, reaching a rate of ≈ 25%.
- Compared to IsolationForest, the OCSVM is much more sensitive to differences in the **initialfootcontact** phase.

4 Classification (Predicting Gait Condition)

1. **Target and Group Definition:** A binary target is_braced is created (1 if condition $\in \{2,3\}$, 0 if condition = 1). The subject IDs (groups) are used to define the cross-validation folds, making the model generalize to new subjects.

$$y = \text{df['is\_braced']}, \quad \text{Groups} = \text{df['subject']}$$

2. **Group K-Fold Setup:** A GroupKFold strategy with $n\_splits = 8$ is initialized. This guarantees that all data samples belonging to the same subject are kept together in either the training or testing set for any given fold.

$$\text{GroupKFold(n\_splits} = 8)$$

3. **Model Pipeline and Training:** A Pipeline is used within each fold to ensure correct sequential processing: first, StandardScaler is fit only on the training data, and then the model (clf) is trained.

$$\text{Pipeline} = [\text{StandardScaler}(), \text{Classifier}(\cdot)]$$

4. **Iterative Evaluation:** The three models (LogisticRegression, RandomForest, GradientBoosting) are evaluated across all 8 subject-split folds. Performance is measured using Accuracy, F1-score, and AUC (Area Under the ROC Curve).

Model Performance Summary

The table below summarizes the mean performance across the 8 folds, providing a reliable measure of each model's ability to generalize to unseen subjects.

Table 1: Performance Metrics Across 8 Group K-Folds (Mean Values)

Model	**Metric**	**ACC Mean**	**F1 Mean**	**AUC Mean**
RandomForest	Mean	**0.725**	**0.805**	**0.778**
GradientBoosting	Mean	0.709	0.796	0.759
LogReg	Mean	0.603	0.624	0.707

Key Findings from Cross-Validation

- The RandomForest classifier is the best performing model, achieving the highest average metrics: Accuracy(≈ 0.725), F1-score(≈ 0.805), and AUC(≈ 0.778).
- The LogisticRegression performs significantly worse than the ensemble methods, confirming that the relationship between the raw sample-wise features and the gait condition is non-linear.
- **Conclusion on Efficacy:** other than groupKfold girdsearch would be expected to be implemented to get us better results but since it only gave a 2 to 5 % jump from the existing 72% accuracy from the randomforest model from GroupKfold the conclusion wont change
- **Overall Conclusion:** The micro-model features, which use raw sample-wise data, are not effective at reliably classifying gait anomalies (braced vs. unbraced) when generalizing across different subjects.

Combined Summary of Gait Analysis

Feature Engineering:

- Macro-level features (velocity, acceleration, range, CV, smoothness) effectively capture overall movement variability and control across gait cycles.
- Micro-level features (jerk, symmetry index, stability score, gait score) reveal fine-grained joint behavior and coordination at each timepoint.
- Macro features are better for classification and anomaly detection; micro features are richer for biomechanical interpretation.

EDA(Condition and Joint-Level Pattern):

- Condition 1 (Unbraced) shows highest angle means and widest variability.
- Condition 3 (Ankle Braced) consistently lowers angle mean and tightens distributions, especially in the hip.
- Knee joint is most sensitive to bracing, especially under knee brace (Condition 2).
- The ankle joint moves the most suddenly and unevenly, while hip remains the most stable.
- Bracing affects joint dynamics different ways, knee brace alters range and smoothness, ankle brace shifts timing and coordination.

Anomaly Detection:

- Isolation Forest trained on healthy data isolates braced cycles with lower anomaly scores.
- One-Class SVM is more sensitive to anomalies at the start of the gait cycle (0%) and shows sharper separation.

Classification:

Macro-Level

- Logistic Regression achieves 95.6% accuracy with strong linear separation.
- Random Forest captures non-linear interactions (88.9% accuracy).

Micro-Level

- Random Forest generalizes best across subjects (Accuracy: 72.5%, F1: 80.5%, AUC: 77.8).
- Gradient Boosting performs similarly but slightly lower.
- Logistic Regression underperforms due to non-linear feature relationships.

This analysis highlights how bracing alters gait dynamics in measurable ways—reducing natural motion and coordination. Combining macro-level features with anomaly scores offers a reliable framework for assessing gait health across subjects.