Interpretable machine learning for classifying time-defined post-injury recovery stage from gait data after spinal cord injury

1. Abstract

Background: CatWalk XT is a tool for objective analysis of motor function in rats. The numerous, often correlated, and mostly biologically opaque raw parameters produced by it make assessing the motor function recovery in rats after a thoracic spinal cord injury (tSCI) difficult. Deriving a subset of these parameters correlated to time-defined recovery still presents a major challenge for researchers testing efficacy of novel treatments in rats after tSCI.

Objective: The objectives were to train and compare XGBoost, LightGBM, and CatBoost classifiers using engineered CatWalk XT features for time-defined recovery-phase classification, and then to combine the three classifiers' three-feature outputs into a continuous equal-weight ensemble recovery index.

Methods: The initial dataset comprised 396 records from 115 Wistar rats. Animals were subjected to compression/contusion clip injury at the level of Th9. CatWalk XT trials were performed on postoperative days 3, 7, 14, 21, 28, and 35. 26 engineered features representing different physical domains were calculated from the raw run parameters. Based on the post-op day of measurement, trials were assigned to early (days 3 and 7), middle (days 14 and 21), or late (days 28 and 35) recovery phases. XGBoost, LightGBM, and CatBoost classifiers were trained and evaluated using the engineered-feature dataset. SHAP analysis and feature ablation were used to interpret the classifiers and identify a consensus panel comprising the three highest-ranked features. The classifiers were subsequently retrained using this panel, and their outputs were combined to derive a continuous ensemble recovery index. Uncertainty in the mean recovery index was estimated using 5,000 experiment-stratified bootstrap resamples at the animal level.

Results: CatBoost, LightGBM, and XGBoost achieved 26-feature mean cumulative AUROCs of 0.861 (95% CI, 0.832–0.889), 0.845 (95% CI, 0.811–0.878), and 0.835 (95% CI, 0.801–0.869), respectively. Forepaw intensity density ranked highest in the consensus feature-importance analysis and produced the largest mean decrease in AUROC when ablated. After retraining on the three-feature panel, the respective mean AUROCs were 0.856 (95% CI, 0.826–0.886), 0.856 (95% CI, 0.826–0.886), and 0.859 (95% CI, 0.830–0.888). The ensemble recovery index increased across predefined phases, from 24.1 (95% CI, 21.9–26.7) in Early trials to 49.3 (95% CI, 44.1–54.4) in Middle trials and 63.9 (95% CI, 58.8–68.8) in Late trials.

Conclusion: A machine-learning workflow based on 26 features engineered from CatWalk XT parameters discriminated among time-defined recovery phases after tSCI, with consistently good internal performance across CatBoost, LightGBM, and XGBoost. Forepaw intensity density was the most consistently important predictive feature, while the three-feature classifier ensemble yielded a continuous index that increased across the predefined recovery phases.

Keywords: spinal cord injury; CatWalk XT; gait analysis; machine learning; SHAP; recovery index.

2. Introduction

Thoracic spinal cord injury at T9 in rats commonly impairs hindlimb motor function and somatosensory pathway transmission [1, 2]. Urinary and bowel dysfunction can also occur [3, 4]. An analysis of the movement patterns arising as a result of hindlimb function loss presents a challenge because motor recovery in rats after tSCI is assessed with complementary behavioral and gait-based measures rather than a single definitive test [5–8]. The heterogeneous, often non-linear post-injury changes in diverse gait parameters make defining a singular index for quantitatively evaluating the gradual recovery of motor function very difficult [7–14].

One of the most popular tools for evaluating the hindlimb function after an SCI in rats is the Basso, Beattie, and Bresnahan scale [5]. Although it is widely used and reliable, it does not produce parameters, that objectively evaluate different facets of hindlimb motor function on a continuous scale [6, 8]. CatWalk XT is an automated gait measurement and analysis system, which utilizes the behaviour of the reflected light to detect, classify, and analyze individual paw prints [6, 7, 14–16]. These measurements are used to produce a large set of raw features [6–8, 13–19], spanning from geometrical parameters, such as contact area for each individual paw, to temporal parameters, such as the duration of the mean relative speed of a paw to the body during a swing, to intensity parameters, such as the maximal and minimal paw contact intensity during a step cycle.

The data, produced by the CatWalk XT in a given trial, contains a high number of very specific objective measures of paw motor function. These parameters are often highly correlated, sparse, and biologically opaque, and as such it is rather difficult for a researcher to identify a subset of parameters, that should be used to assess the phase or trajectory of motor function recovery [13, 14, 20]. The analysis of the measurements requires a higher level of statistical rigor because of the additional challenge of repeated measurements within individual animals at separate timepoints and the associated confounding effects [21].

Gradient-Boosted Decision Trees suit the task of analyzing CatWalk XT data well: they natively possess the ability to handle data sparsity, are able of learning non-linear relationships between the data and the target, and are stable in regard to the input feature value scaling [22–26]. As is the case with many advanced machine-learning approaches, this classifier type is often treated as a black box, so post hoc methods are needed to assess how individual features contribute to a prediction [27–29]. In order to identify the features, that have the biggest influence on model accuracy, there exist at least two methodically different, but complementary approaches.

The first widely used method is leave-one-feature-out ablation, which assesses the drop in performance of a given model on a predefined constant dataset, when a singular feature is omitted from analysis [20, 29, 30]. This retraining-based procedure differs from permutation importance, which perturbs feature values without omitting the predictor from model fitting [30, 31]. The resulting ranked list of changes in a selected score metric is then to be considered a proxy for feature importance from most to least (or vice-versa). Importantly, this method does not deliver any information on how exactly a value of a given feature influences the model prediction [29, 30].

The second is a game-theoretical approach utilizing Shapley values. This method estimates the magnitude and direction of each feature's attribution to the model output for each trial [27, 28]. By calculating the statistics for the entirety of computed Shapley values, it is possible to create a feature table ranked by importance for a selected classifier type and dataset.

By combining these approaches, a subset of consensus high-influence features can be identified by comparing the feature lists across models and approaches. From the resulting subset, an n-feature panel can be selected to be used to retrain a new ensemble of classifiers of the same type. Subsequently, the significance of performance

discrepancy can be calculated to decide, if the reduced-feature dataset allows for acceptable representation of the underlying data signal while lowering its dimensionality to the selected n [20, 26].

Assuming the adequate performance of the classifiers using the reduced-feature dataset, an ensemble recovery index can be derived as a continuous proxy for a discrete ordinal classification of early-to-middle-to-late recovery phase [32, 33].

In this study, a combined dataset of CatWalk XT trials from five previous tSCI experiments was analyzed utilizing the described approach.

3. Materials and methods

3.1 Study design, source experiments, and animals

This paper shows a post-acquisition analysis with no new intervention performed. The five source experiments were completed in the same lab on Wistar rats from the same supplier. The CatWalk XT hardware and software suite did not differ in version between the experiments.

In all source experiments, the rats were subjected to a comparable surgical intervention performed by three researchers. To induce a tSCI, a laminectomy with subsequent T9 clip contusion/compression for 60 seconds using a 28-g modified aneurysm clip was performed. This procedure is based on the established aneurysm-clip compression model [34]. The surgery was performed under isoflurane anaesthesia. After the operation rats received analgesia, as well as antibiotic and saline support. Manual bladder voiding was regularly performed until recovery of bladder control. The age of the animals ranged from 6 to 8 weeks at the start of the experiment. The body weight of the animals during the experiments was approximately 230-265 g. All rats were group housed, and underwent seven days of CatWalk acclimation. G-211/15 under German animal-protection requirements.

An equivalent CatWalk XT setup was used to acquire trials at days 3, 7, 14, 21, 28, and 35 post-injury. CatWalk XT acquisition and parameter conventions have been described previously [6, 7, 14–16]. The acquisition of CatWalk XT data was performed by four researchers. A trial was to be deemed valid if the animal completed three full runs through the detection area. Subsequently, built-in automatic paw print classification was performed and then corrected and validated by three independent researchers. Mean values for each CatWalk XT parameter were calculated and selected as being representative of a given trial.

3.2 Analytical populations and ordinal outcome

The initial combined five-experiment dataset contained 2,101 rows from 451 animals. After quality control and exclusion as specified in Supplementary Table S2, 1,524 rows from 387 animals were retained. The raw CatWalk XT parameters aggregated corresponding left and right paw measurements where applicable. Additionally, the corresponding recovery phase was defined for each trial, depending on the post-operative day of data acquisition: 3rd and 7th days were defined as early, 14th and 21st days as middle, and 28th and 35th days as late recovery stage.

Only animal which have undergone a complete surgical intervention, but received no treatment. were included in the analysis cohort. These corresponded to being the control group in the source experiments. After selection, the resulting dataset consisted of 396 recordings from 115 animals. Most animals contributed recordings to more than one stage, so stage-specific animal counts are not additive. These recordings exclusively were used for model training, feature importance calculation, and recovery index derivation.

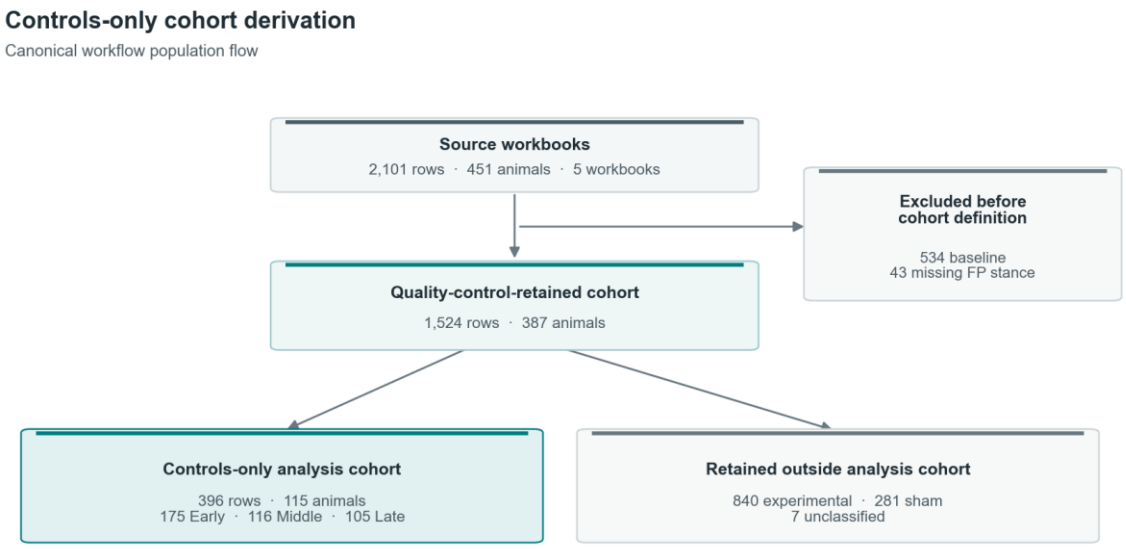

Figure 1. Derivation of the controls-only population.

Table 1. Full and controls-only populations

Population	Records	Animals	Analytical role
All source experiments	2,101	451	-
After quality control	1,524	387	Exclusion criteria specified in Supplementary Table S2.
Controls-only cohort	396	115	Early 175 (93 animals); Middle 116 (58); Late 105 (33).

3.3 Harmonization and engineered features

The compression/contusion clip injury was laterally symmetrical, and thus the motor regeneration was also assumed to be symmetrical. Considering that, the values for right and left paws for each corresponding parameter were aggregated into a singular frontpaw or hindpaw parameter, calculated as a simple arithmetic mean.

A predefined set of basic formulas was used to calculate 26 engineered variables spanning peak loading, loading fraction, speed-normalized intensity, functional output, integrated loading, loading rate, propulsion, symmetry, temporal contact and stride-per-cycle domains. The exact specifications for this calculation are described in Supplementary Table S5. All classifiers were trained exclusively on the calculated engineered features. The selected three-feature panel also consisted only of engineered features.

Table 2. Post-selection three-feature consensus panel

Feature	Formula	Role in consensus
fp_intensity_density	FP intensity / (FP maximum contact area + ε)	Combined consensus rank 1; strongest SHAP and ablation evidence.
fp_intensity_loading_fraction	FP intensity / (FP intensity + HP intensity + ε)	Combined consensus rank 2; cross-compartment loading fraction.
hp_loading_rate	HP intensity / (HP stance + ε)	Combined consensus rank 3; complementary hindpaw loading rate.

Table note. FP and HP are bilateral forepaw and hindpaw compartments. The three-feature panel was selected after analysis of the 26-feature OOF evidence and was evaluated as a post-selection sensitivity analysis before construction of the ensemble recovery index.

3.4 Missing data

The entirety of the missing data represents an absent hindpaw detection as a result of missing hindlimb motor function. Missing values were not modified or imputed in any way before the analysis and were handled natively by the classifiers. Across the entire dataset of 396 records, after calculating engineered features, 4,179 of 10,296 cells were missing (40.6%), and 113 of 396 records were complete. Missingness was 56.6% in Early, 32.7% in Middle, and 22.6% in Late. The three-feature panel contained 452 missing values among 1,188 cells (38.0%), and 162 of 396 records were complete across all three features.

3.5 Ordinal models, feature selection, and nested validation

For each of the classifiers, two cumulative binary tasks were defined: q1 = P(Y > Early) and q2 = P(Y > Middle) = P(Late) [32, 33]. Ordinal consistency was enforced by replacing q2 with min(q2, q1). Class probabilities were reconstructed as P(Early) = 1 - q1, P(Middle) = q1 - q2, and P(Late) = q2. The largest reconstructed probability was used to evaluate hard classification accuracy.

The outer folds were generated using the recovery phase for stratification and the animal identificator as the grouping variable [21]. Outer test folds contained 79, 79, 78, 80, and 80 records. For each outer fold, a five-fold stratified grouped inner validation was performed [21, 35]. For every framework-by-outer-fold combination, the classifier hyperparameters were optimized in an independent 100-trial Optuna study, with mean Early/Late AUROC as the metric to be maximized [36].

After hyperparameter optimization, all 26 engineered features were used for the evaluation of the classifiers. During SHAP analysis and feature ablation [20, 27–30], the same predefined outer folds were used. The subsequently selected three-feature panel underwent a separate nested cross-validated rerun with the same frozen outer-fold structure [21, 35].

3.6 Performance, calibration, interpretation, and sensitivity analyses

Recovery phase classification was defined as an ordinal problem, with the three phases ordered from Early to Middle to Late recovery. Two cumulative binary classification tasks were evaluated. The q1 probability was used to calculate the AUROC for distinguishing Early recovery from Middle or Late recovery, whereas q2 was used to calculate the Late AUROC by distinguishing Early or Middle recovery from Late recovery. Mean AUROC was calculated as the simple arithmetic mean of these two values [37, 38]. To obtain the pooled out-of-fold estimates, predictions were calculated for the test partition of each outer fold and subsequently concatenated. The corresponding 95% confidence intervals and paired differences between the evaluated frameworks were estimated using 5,000 bootstrap replicates clustered by animal [39, 40]. Bootstrap sign-tail p values were considered descriptive and were not adjusted for multiple comparisons.

Different calibration scenarios, such as no calibration, logistic (Platt) calibration, and isotonic calibration, were evaluated for q1 and q2 [41–43]. The calibration models were fitted using animal-grouped inner out-of-fold predictions and subsequently applied to the corresponding untouched outer-test partitions [21, 35, 44, 45]. A conservative selection rule was used, according to which calibration was selected only if the clustered 95%

confidence interval for improvement in the mean cumulative Brier score excluded zero and the pooled log loss did not worsen [38, 41, 42, 46]. Based on these criteria, logistic calibration was selected for the XGBoost and LightGBM classifiers, whereas no calibration was selected for CatBoost. Consequently, CatBoost remained uncalibrated in this framework-specific calibration analysis.

Prediction-level robustness was evaluated across row-weighted pooled, equal-animal, and equal-outer-fold estimands, under uncalibrated and framework-selected calibration scenarios. Each estimand was assessed for both the uncalibrated predictions and the predictions obtained after applying the previously selected framework-specific calibration.

3.7 Ensemble recovery index and additive surrogate

For each of the three resulting classifiers utilizing the same three-feature panel, an explicitly defined recovery-index was calculated:

$$RI = 50 \times P(\text{Middle}) + 100 \times P(\text{Late}).$$

From the framework-specific indices the equal-weight mean was calculated to form an equal-weight ensemble recovery index. Each index ranges from 0 to 100 and represents expected ordinal class position, not percentage of biological recovery. To simplify and parameterize the calculation of the recovery index, an additive-spline surrogate was fitted [26, 47]. The generalized additive model was evaluated for fidelity to the model-derived index; it was not an independently validated classifier of observed recovery stage.

3.8 Software and reporting

Analyses used Python 3.10.8, scikit-learn 1.7.2, XGBoost 3.2.0, LightGBM 4.7.0, CatBoost 1.2.10, Optuna 4.9.0 [24, 25, 36, 48, 49], pandas 2.3.3, NumPy 2.2.6, SciPy 1.15.3, and random seed 0. Workflow 09 used native tree-library SHAP interfaces and did not require the separate shap package [27, 28]. Reporting and risk-of-bias considerations were informed by ARRIVE 2.0, TRIPOD+AI, and PROBAST+AI [50–53].

4. Results

4.1 Cohort composition and missingness

The full dataset consisting of all records from five source experiments contained 2,101 records. The controls-only cohort comprised 396 records from 115 control animals: 175 Early records from 93 animals, 116 Middle records from 58 animals, and 105 Late records from 33 animals. All reported classifier performance estimates were based on cross-fitted predictions, with observations grouped by animal (Figure 1; Table 1; Supplementary Tables S2–S3).

Missingness was heavily stage dependent and represented a lack of hindlimb function. Of 10,296 engineered values, 4,179 (40.6%) were missing; 113 records (28.5%) were complete across all 26 features. Early, Middle, and Late missingness proportions were 56.6%, 32.7%, and 22.6%, respectively. Tree models retained all eligible rows and used native missing-value handling (Supplementary Table S6).

4.2 Discrimination and hard classification

All three 26-feature classifiers showed useful pooled out-of-fold discrimination (Table 3; Figures 2–3). CatBoost achieved Early, Late, and mean AUROCs of 0.871, 0.852, and 0.861. LightGBM achieved 0.859, 0.830, and 0.845; XGBoost achieved 0.852, 0.817, and 0.835. The animal-clustered 95% CI for CatBoost mean AUROC was 0.832–0.889. The CatBoost–XGBoost mean-AUROC difference was 0.026 (95% CI, 0.009–0.045); the CatBoost–LightGBM difference was 0.016 (95% CI, −0.002 to 0.035), and the LightGBM–XGBoost difference was 0.010 (95% CI, −0.006 to 0.026). Thus, only the CatBoost–XGBoost interval excluded zero in the 26-feature comparison. This comparison did not define a single-framework application artifact; the

later recovery-index workflow combined the fixed three-feature outputs by prespecified equal weighting (Supplementary Tables S9–S10; Supplementary Figure S1).

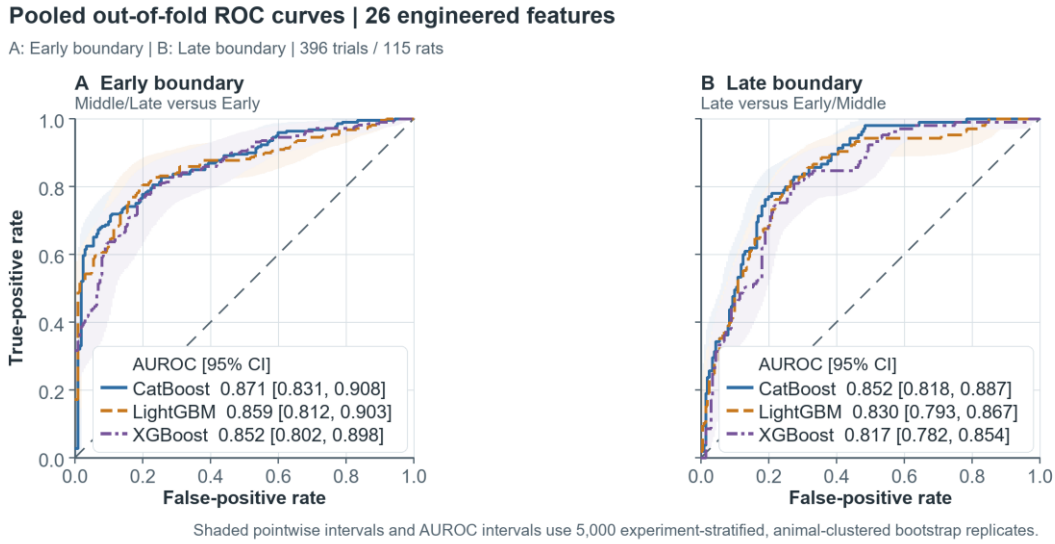

Figure 2. Pooled out-of-fold receiver operating characteristic curves for the 26-feature classifiers. (A) Early cumulative boundary (Middle/Late versus Early). (B) Late cumulative boundary (Late versus Early/Middle).

For each classifier, curves use one pooled outer-test prediction for each of 396 records from 115 rats. Each record was predicted by a model that had not been trained on that animal. Legends report pooled AUROC with experiment-stratified, animal-clustered 95% confidence intervals; shaded bands show pointwise 95% intervals. The dashed diagonal denotes no discrimination.

Table 3. Pooled out-of-fold discrimination across the three 26-feature classifiers

Framework	Early AUROC [95% CI]	Late AUROC [95% CI]	Mean AUROC [95% CI]	Accuracy / balanced accuracy
CatBoost	0.871 [0.831-0.908]	0.852 [0.818-0.887]	0.861 [0.832-0.889]	0.644 / 0.606
LightGBM	0.859 [0.812-0.903]	0.830 [0.793-0.867]	0.845 [0.811-0.878]	0.636 / 0.596
XGBoost	0.852 [0.802-0.898]	0.817 [0.782-0.854]	0.835 [0.801-0.869]	0.621 / 0.580

Table note. Estimates use all 396 mutually exclusive outer-test records. Confidence intervals are from 5,000 animal-clustered bootstrap replicates stratified by experiment.

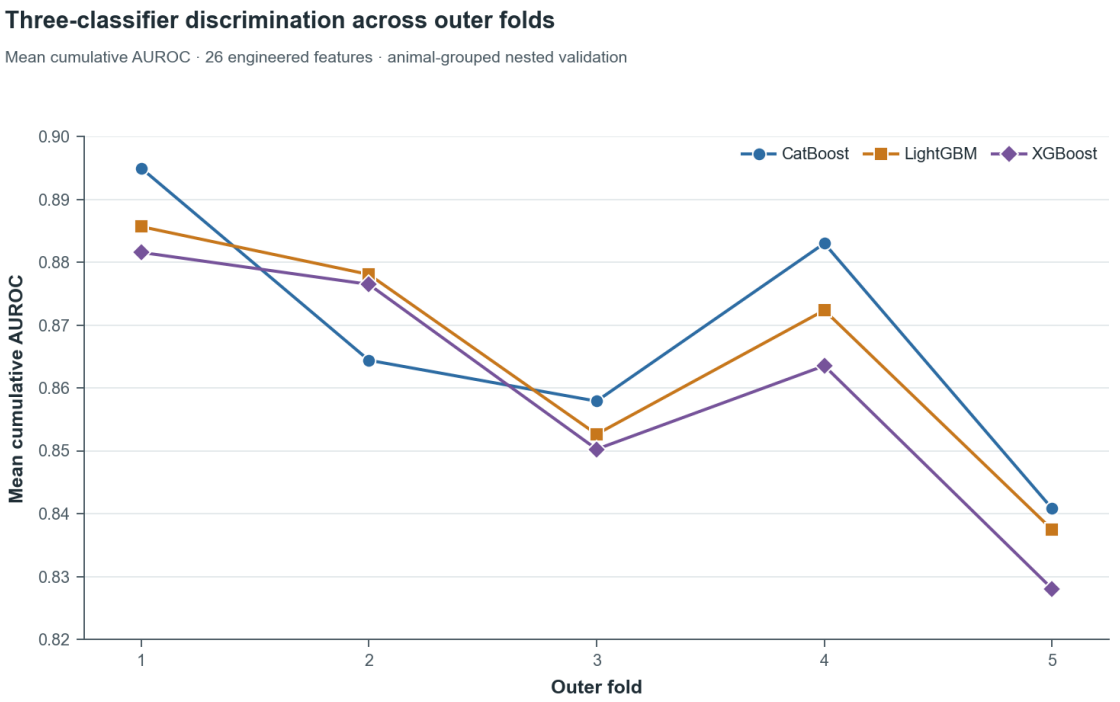

Figure 3. Three-classifier discrimination across animal-grouped outer folds.

Lines and points show mean cumulative AUROC by outer fold for the three 26-feature classifiers. Between-fold variation was larger than most between-framework differences.

Hard classification was most reliable for Early observations and least reliable for Middle observations. Early recall ranged from 85.1% to 86.3%, Middle recall from 32.8% to 42.2%, and Late recall from 54.3% to 58.1%. Accuracy, balanced accuracy, and ordinal MAE are reported together for all three frameworks in Supplementary Table S9. Most errors were adjacent-stage assignments (Figure 4; Supplementary Figure S3).

Calibration effects were framework dependent. Logistic calibration improved mean cumulative Brier score for XGBoost by 0.0081 (95% CI 0.0022-0.0147) and for LightGBM by 0.0065 (95% CI 0.0007-0.0126). CatBoost logistic improvement was 0.0012 (95% CI -0.0015 to 0.0040), while logistic loss worsened; therefore, CatBoost remained uncalibrated under this framework-specific rule (Supplementary Table S11; Supplementary Figure S2).

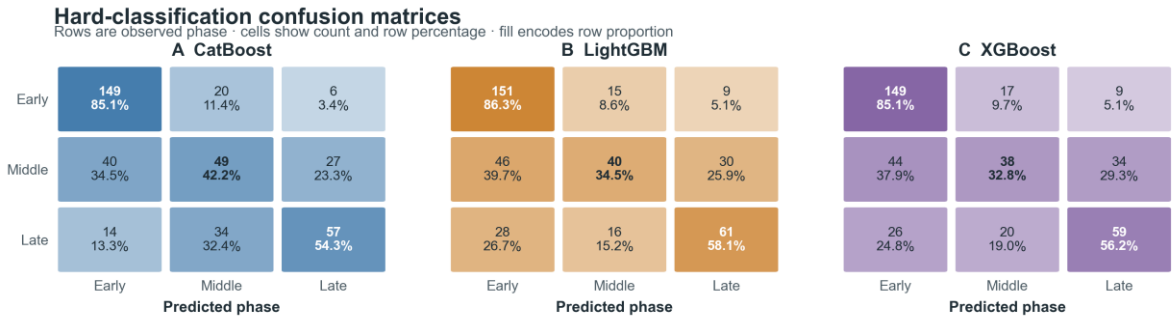

Figure 4. Raw out-of-fold hard-classification confusion matrices.

Cells report counts and percentages within observed-stage rows for the three 26-feature classifiers.

4.3 Feature selection and post-selection three-feature sensitivity analysis

4.3.1 Native SHAP and feature ablation identified a dominant feature

Native SHAP consistently ranked fp_intensity_density first across the three classifiers. SHAP consensus between the frameworks next prioritized fp_intensity_loading_fraction, fp_loading_speed, hp_loading_rate, and fp_temporal_contact (Figure 5; Supplementary Figures S4–S7).

Paired leave-one-feature-out ablation also identified fp_intensity_density as the most important feature, defined as the feature whose removal led to the highest mean AUROC loss. Removing it reduced mean AUROC by an average of 0.0224 across frameworks; all three framework-specific unadjusted clustered intervals excluded zero, although none remained significant after global Holm adjustment. Other feature-removal effects were smaller and less consistent, compatible with redundancy among correlated engineered variables (Figure 6; Supplementary Figure S8).

Agreement among the 15 framework-by-outer-fold SHAP rankings was stronger (Kendall W = 0.686) than among the 15 matched ablation rankings (W = 0.168). SHAP-consensus versus ablation-consensus ranks showed modest concordance (Spearman $\rho = 0.217$; Kendall $\tau_b = 0.169$; rank R-squared = 0.047). The equal-weight 30-judge consensus therefore integrated complementary evidence rather than assuming equivalent rankings (Supplementary Table S14; Supplementary Figures S9–S10).

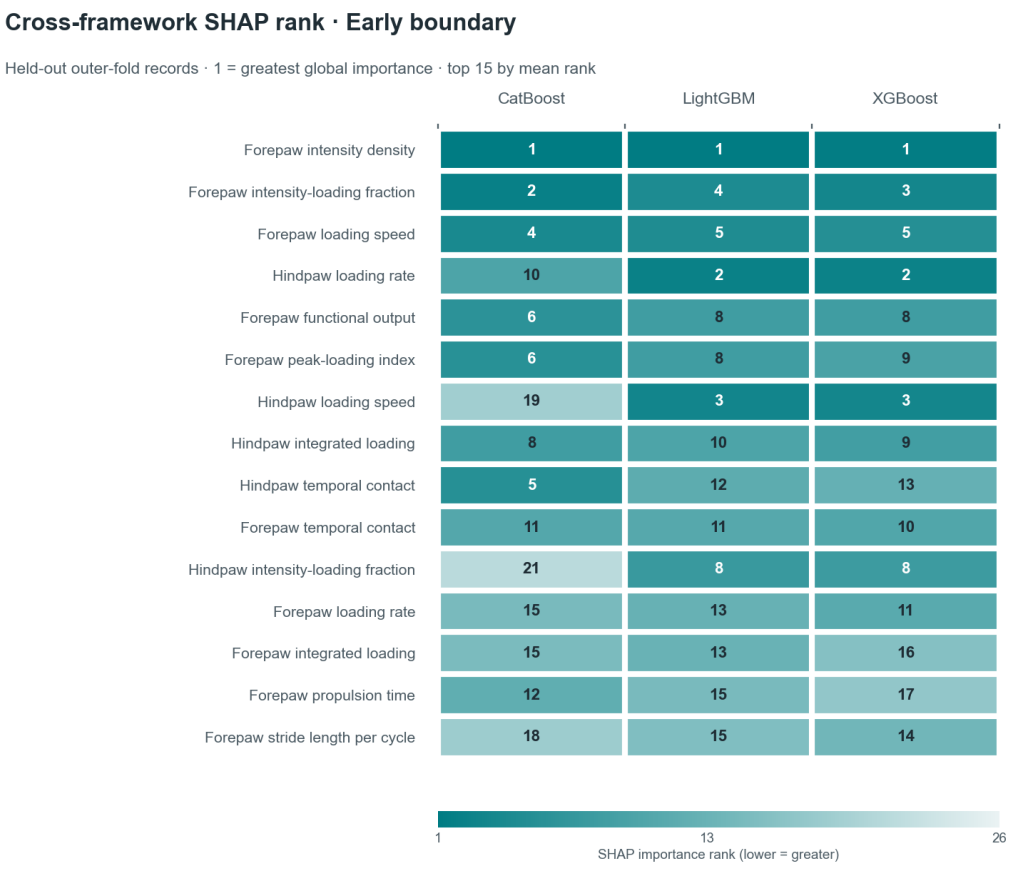

Figure 5. Cross-framework SHAP importance ranks for the Early cumulative threshold.

Cells report held-out SHAP importance ranks (1 = greatest) from the three 26-feature outer-fold classifiers. Ranks permit cross-framework comparison; SHAP does not establish causal importance[27–29].

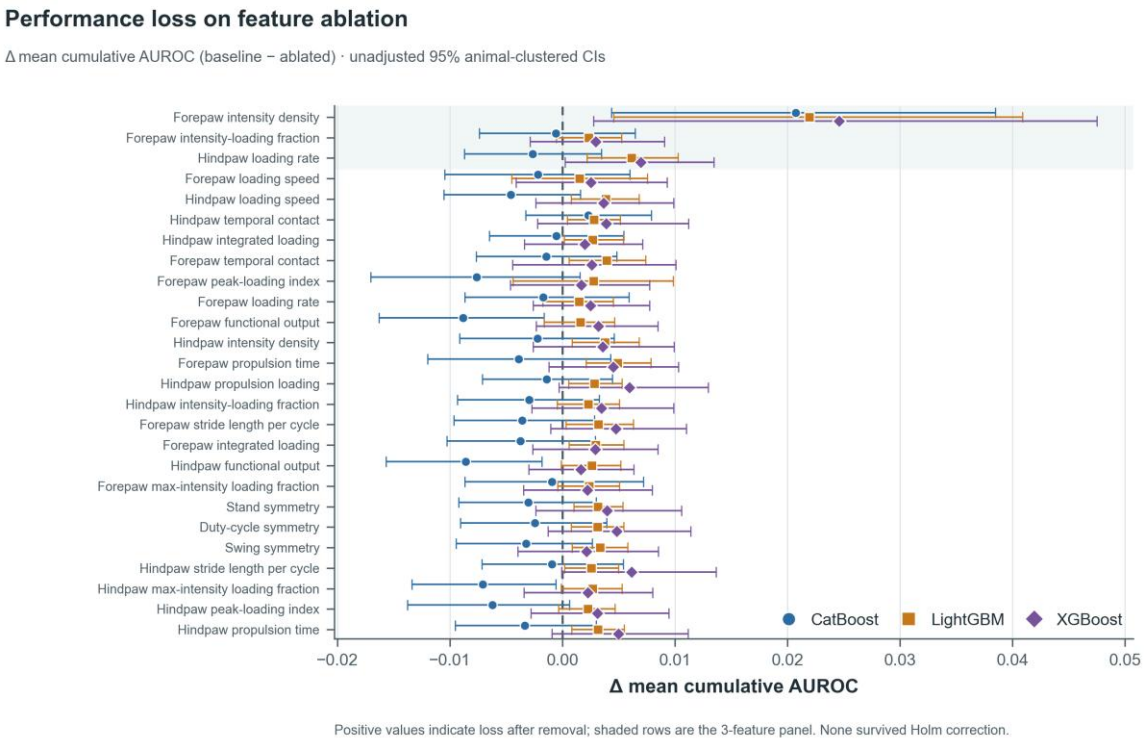

Figure 6. Paired leave-one-feature-out mean-AUROC effects across the three 26-feature classifiers.

Positive values indicate worse held-out discrimination after removal. Intervals are unadjusted animal-clustered 95% CIs; none remained significant after Holm correction across 26 features.

4.3.2 Post-selection three-feature sensitivity analysis

Within each of 15 framework-by-outer-fold analyses, features were ranked separately by held-out SHAP importance and held-out mean-cumulative-AUROC loss after ablation. The resulting 30 ranks were pooled with equal weight using a mean-rank (Borda) consensus; ties were resolved by lower median rank, lower rank SD, then feature name. Evidence tiers and alternative method weightings were descriptive and did not determine selection. The first three features were fp_intensity_density, fp_intensity_loading_fraction, and hp_loading_rate. The resulting three-feature dataset was used to train, optimize, and evaluate a new set of classifiers on the same outer-fold structure. Because selection used 26-feature out-of-fold evidence from the same 115 animals, the fixed-panel rerun was interpreted as post-selection sensitivity rather than independent validation (Table 2; Supplementary Table S13).

The three-feature pooled mean AUROCs were 0.859 (95% CI, 0.830–0.888) for XGBoost, 0.856 (95% CI, 0.826–0.886) for CatBoost, and 0.856 (95% CI, 0.826–0.886) for LightGBM, compared with the corresponding 26-feature estimates of 0.835 (95% CI, 0.801–0.869), 0.861 (95% CI, 0.832–0.889), and 0.845 (95% CI, 0.811–0.878), respectively. The descriptive changes were +0.025, −0.005, and +0.011. This same-cohort comparison suggests that the consensus panel retained much of the observed discrimination, but it cannot establish that feature reduction improved generalization (Figure 7; Supplementary Table S15).

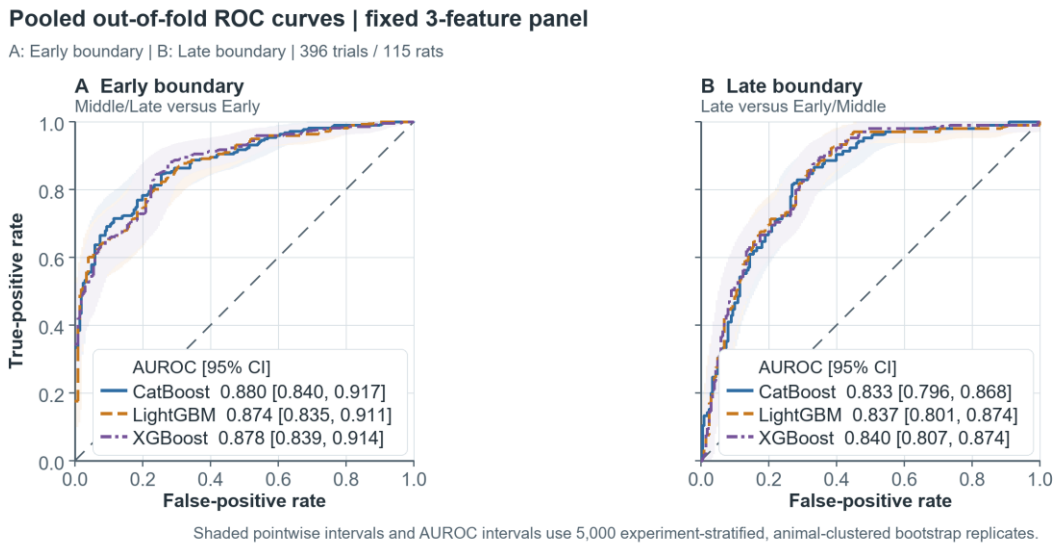

Figure 7. Pooled out-of-fold receiver operating characteristic curves for the post-selection three-feature classifiers. (A) Early cumulative boundary (Middle/Late versus Early). (B) Late cumulative boundary (Late versus Early/Middle).

For each classifier, curves use one pooled outer-test prediction for each of 396 records from 115 rats. Each record was predicted by a model that had not been trained on that animal. Legends report pooled AUROC with experiment-stratified, animal-clustered 95% confidence intervals; shaded bands show pointwise 95% intervals. The dashed diagonal denotes no discrimination. The fixed panel was selected from 26-feature out-of-fold evidence from the same animals, so this is a post-selection sensitivity analysis, not independent validation.

The three-feature calibration analysis selected logistic (Platt) calibration for CatBoost, whereas no calibration was selected for LightGBM or XGBoost under the conservative probability-accuracy rule. For construction of comparable recovery-index components, Workflow 13 nevertheless applied Platt calibration uniformly to all three frameworks. This was an index-construction step rather than a framework-ranking step.

4.4 Ensemble recovery index and GAM surrogate

The equal-weight ensemble recovery index combined Platt-calibrated three-feature CatBoost, LightGBM, and XGBoost probabilities. Record-level means were 24.1 ± 15.9 for Early, 49.3 ± 21.8 for Middle, and 63.9 ± 19.9 for Late. The corresponding medians were 17.0, 47.2, and 70.4. The index was associated with ordered phase (Spearman $\rho = 0.655$) (Figure 8; Supplementary Table S16; Supplementary Figures S11–S12).

The GAM surrogate approximated the ensemble index with out-of-fold $R^2 = 0.957$, RMSE = 5.26, MAE = 4.01, and Spearman $\rho = 0.969$. Its agreement slope was 0.984 and concordance correlation coefficient was 0.978. These are fidelity measures of the additive model surrogate, not independent stage-classification validation. The overlapping index distributions, especially in Middle observations, reinforce that the score is a probability-weighted description rather than a validated biological recovery scale (Supplementary Table S17; Supplementary Figures S13-S15).

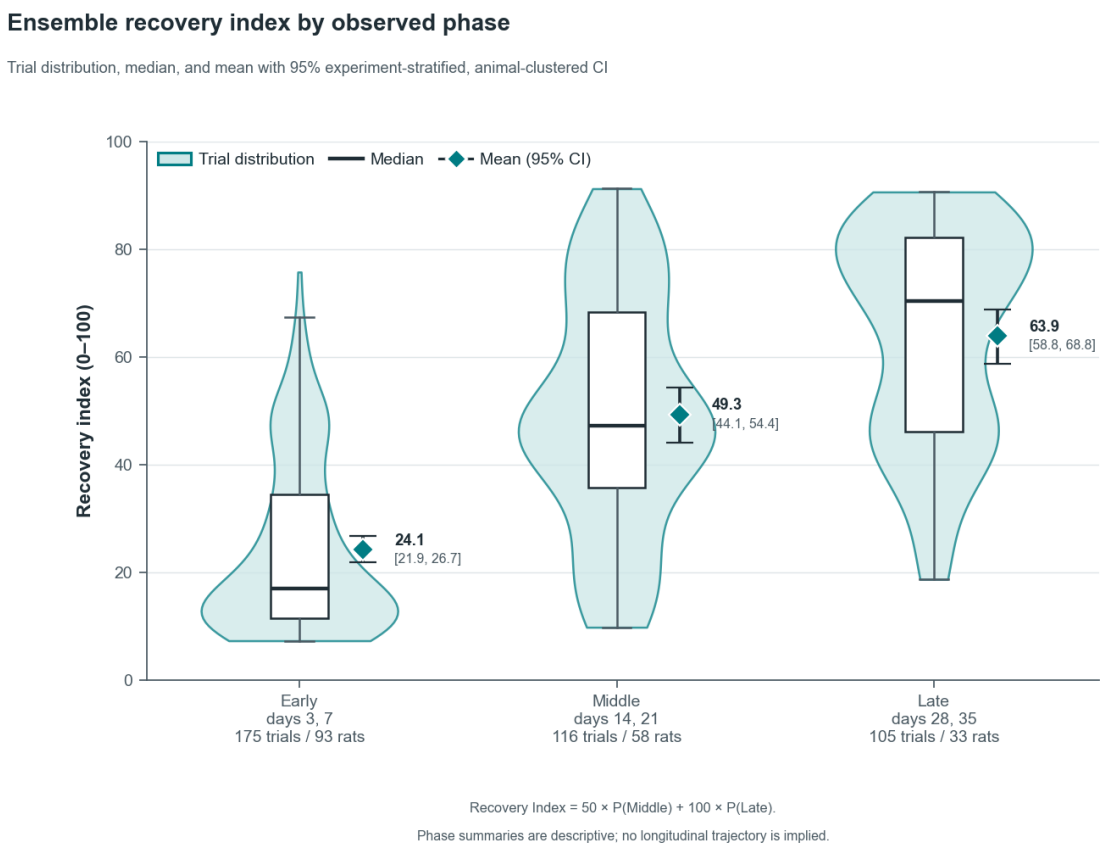

Figure 8. Equal-weight ensemble recovery index by observed phase.

Violins show trial distributions; boxes show medians and interquartile ranges; diamonds and whiskers show means and 95% clustered CIs. The 3-feature ensemble score is not percentage of biological recovery.

5. Discussion

This study provides a framework for classifying the post-tSCI recovery phase from automated gait data. Across three classifier implementations using the same 26 features, all models showed useful discrimination under animal-grouped nested validation; one of three pairwise intervals favored CatBoost over XGBoost, while the other two included zero. All pairwise intervals among the fixed three-feature classifiers included zero. The three-feature analysis retained similar discrimination, and the ensemble recovery index provided a continuous prespecified summary of their ordinal probabilities.

The clearest feature-level finding concerned forepaw intensity density, defined as contact intensity relative to maximum contact area. It ranked first in the SHAP analyses across all three frameworks and produced the largest and most consistent loss in performance during ablation. This suggest, that the forepaw intensity density is a stable predictive signal within a given cohort. It does not, however, imply biological causality or establish that the feature reflects a causal gait mechanism [27–29].

Other loading variables contributed complementary information. SHAP rankings were relatively stable across framework-by-fold analyses, whereas ablation rankings and SHAP–ablation concordance were more modest. This pattern is consistent with redundancy and substitutability among correlated gait predictors [20, 29, 30]. The consensus three-feature panel should therefore be viewed as a pragmatic integration of attribution, functional removal, and cross-framework evidence, rather than as evidence that three independent biomarkers were identified.

The Middle stage was the least separable class. This was expected because the outcome was ordinal and defined using adjacent time windows. Middle probabilities were calculated as the difference between two cumulative

probabilities [32, 33], and the phenotype at days 14–21 may overlap both earlier impairment and later compensation. The confusion matrices therefore showed errors in both directions. Direct Early-to-Late errors were less frequent, supporting the ordinal decomposition, even though three-class accuracy remained moderate. The recovery index represented this overlap more naturally by allowing intermediate probability-weighted values rather than forcing each recording into a discrete stage.

The recovery index is best interpreted as a calibrated ensemble model score. Its ordered distribution across phases provides internal face validity for the learned ordinal representation. The high out-of-fold fidelity of the GAM further shows that the index can be approximated using smooth additive functions of the three selected predictors [26, 47]. Neither finding validates the index against independent functional, histological, electrophysiological, kinematic, or treatment-response outcomes [38, 44, 45].

Animal grouping prevented repeated records from the same rat from entering both training and held-out partitions. Hyperparameter optimization and fitted feature transformation were restricted to outer-training data [21, 35]. The same outer folds were used for paired framework comparison. Calibration was fitted using inner out-of-fold predictions and then applied to untouched outer-test data [41–45]. SHAP was calculated only for held-out observations, and feature-ablation analyses reused the same train-test partitions.

Missingness was retained as part of the observed data structure. Hindpaw-dependent features were frequently unavailable, particularly in Early-stage records. Native tree-based missing-value handling allowed the complete eligible controls-only cohort to be retained without imputation or modification of missing values. However, stage-dependent missingness may itself carry predictive information and may not transport to laboratories using different acquisition, footprint-acceptance, or quality-control procedures [44, 45, 51, 54].

The main limitations of this study are substantial. This was an internal validation study involving 115 control rats retrospectively pooled from five source experiments. Although the procedures were considered comparable, experiment could not be treated as an external test domain because phase support differed between experiments. Generalization to other laboratories, strains, sexes, injury severities, treatment groups, acquisition systems, or processing pipelines therefore remains unknown [38, 44, 45, 51, 54].

The outcome was defined by experiment-specific timepoint mapping rather than by an independent functional or biological reference. The resulting target should therefore be interpreted as a study-defined temporal stage rather than a directly measured biological state.

Bilateral pooling allowed a recording to be retained when only one paw was observed. Although the intended experimental injury was intended to be symmetrical, it removed left-right asymmetry and may have masked unilateral deficits. Fold-level sample sizes also remained modest [55]. Animal-clustered bootstrap inference accounted for repeated measurements but cannot replace external validation [39, 40, 44, 45]. SHAP and ablation describe predictive associations and model behavior rather than causality [27–30]. The recovery index and GAM effects likewise do not establish biomarker validity.

Future work should validate the three-classifier comparison procedure and the locked three-feature equal-weight ensemble recovery index in an independent control cohort spanning the relevant recovery phases [44, 45, 51, 54]. Acquisition, quality control, feature processing, and outcome mapping should be prespecified [50, 56]. Biological validation should compare model predictions with established behavioral, histological, electrophysiological, and kinematic outcomes. Experiment-blocked validation should be attempted only when held-out experiments provide adequate and comparable support across outcome stages [21].

6. Conclusion

A 26-engineered-feature ordinal machine-learning workflow classified time-defined post-injury stage from CatWalk XT recordings with useful internal discrimination across CatBoost, LightGBM, and XGBoost. Forepaw intensity density was the most reproducible predictive feature, while outputs from the post-selection three-feature classifiers were combined by prespecified equal weighting into a continuous ensemble recovery

index that could be closely approximated by a generalized additive model surrogate. These outputs support interpretable internal staging but do not yet constitute biological recovery measurement, treatment-efficacy assessment, or externally validated deployment.

7. Declarations

Author contributions

M.S., A.Y., and O.A.: measurements, conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, writing - original draft, writing - review and editing, and project administration. X.W. and M.L.: measurements, validation, and data curation. S.K.: project administration. All authors approved the manuscript.

Ethics statement

The source documentation identified animal-study approval G-211/15 under German animal-protection requirements. No new animal intervention, allocation, or outcome acquisition was performed for this secondary computational analysis.

Informed consent

Not applicable.

Funding

This research received no external funding.

Competing interests

The authors declare no competing interests.

Data availability

Analysis code, aggregate result tables, de-identified fold assignments, out-of-fold predictions, calibration and sensitivity outputs, feature-attribution and ablation summaries, recovery-index outputs, and the GAM surrogate package are available through the curated project repository at https://gitlab.com/sergienko.michael/sci. Raw CatWalk workbooks and derived animal-level tables are not publicly distributed. Access requests should be directed to the corresponding author and are subject to institutional approval, applicable ethics requirements, and any required data-use agreement.

Acknowledgements

The authors acknowledge Claudia Pitzer from the Interdisciplinary Neurobehavioural Core Facility at the University of Heidelberg for assistance with neurobehavioural testing.

8. References

1. Cao Q, Zhang YP, Iannotti C, DeVries WH, Xu XM, Shields CB, Whittemore SR. Functional and electrophysiological changes after graded traumatic spinal cord injury in adult rat. Exp Neurol. 2005;191 Suppl 1:S3–S16.
2. Wedekind C, Ullrich R, Klug N. F-wave amplitudes indicate evolving spinal autonomy during spontaneous recovery of hindlimb function in rat spinal cord contusion. Spinal Cord. 2006;44(1):44–48.
3. Hyun JK, Lee YI, Son YJ, Park JS. Serial changes in bladder, locomotion, and levels of neurotrophic factors in rats with spinal cord contusion. J Neurotrauma. 2009;26(10):1773–1782.
4. Willits AB, Kader L, Eller O, Roberts E, Bye B, Strope T, et al. Spinal cord injury-induced neurogenic bowel: a role for host-microbiome interactions in bowel pain and dysfunction. Neurobiol Pain. 2024;15:100156.
5. Basso DM, Beattie MS, Bresnahan JC. A sensitive and reliable locomotor rating scale for open field testing in rats. J Neurotrauma. 1995;12(1):1–21.

6. Hamers FPT, Lankhorst AJ, van Laar TJ, Veldhuis WB, Gispen WH. Automated quantitative gait analysis during overground locomotion in the rat: its application to spinal cord contusion and transection injuries. J Neurotrauma. 2001;18(2):187–201.
7. Hamers FPT, Koopmans GC, Joosten EAJ. CatWalk-assisted gait analysis in the assessment of spinal cord injury. J Neurotrauma. 2006;23(3–4):537–548.
8. Koopmans GC, Deumens R, Honig WMM, Hamers FPT, Steinbusch HWM, Joosten EAJ. The assessment of locomotor function in spinal cord injured rats: the importance of objective analysis of coordination. J Neurotrauma. 2005;22(2):214–225.
9. Ahuja CS, Wilson JR, Nori S, Kotter MRN, Druschel C, Curt A, Fehlings MG. Traumatic spinal cord injury. Nat Rev Dis Primers. 2017;3:17018.
10. Courtine G, Sofroniew MV. Spinal cord repair: advances in biology and technology. Nat Med. 2019;25:898–908.
11. Barrière G, Leblond H, Provencher J, Rossignol S. Prominent role of the spinal central pattern generator in the recovery of locomotion after partial spinal cord injuries. J Neurosci. 2008;28(15):3976–3987.
12. Lankhorst AJ, ter Laak MP, van Laar TJ, van Meeteren NL, de Groot JC, Schrama LH, et al. Effects of enriched housing on functional recovery after spinal cord contusive injury in the adult rat. J Neurotrauma. 2001;18(2):203–215.
13. Timotius IK, Bieler LS, Couillard-Despres S, Sandner B, Garcia-Ovejero D, Labombarda F, et al. Combination of defined CatWalk gait parameters for predictive locomotion recovery in experimental spinal cord injury rat models. eNeuro. 2021;8(2):ENEURO.0497-20.2021.
14. Timotius IK, Roelofs RF, Richmond-Hacham B, Noldus LPJJ, von Hörsten S, Bikovski L. CatWalk XT gait parameters: a review of reported parameters in pre-clinical studies of multiple central nervous system and peripheral nervous system disease models. Front Behav Neurosci. 2023;17:1147784.
15. Chen YJ, Cheng FC, Sheu ML, Su HL, Chen CJ, Sheehan J, Pan HC. Detection of subtle neurological alterations by the CatWalk XT gait analysis system. J Neuroeng Rehabil. 2014;11:62.
16. Heinzel J, Swiadek N, Ashmwe M, Rührnößl A, Oberhauser V, Kolbenschlag J, Hercher D. Automated gait analysis to assess functional recovery in rodents with peripheral nerve or spinal cord contusion injury. J Vis Exp. 2020;(164):e61852.
17. Vrinten DH, Hamers FPT. CatWalk automated quantitative gait analysis as a novel method to assess mechanical allodynia in the rat: a comparison with von Frey testing. Pain. 2003;102(1–2):203–209.
18. Zheng G, Younsi A, Scherer M, Riemann L, Walter J, Skutella T, Unterberg A, Zweckberger K. The CatWalk XT gait analysis is closely correlated with tissue damage after cervical spinal cord injury in rats. Appl Sci. 2021;11(9):4097.
19. Zheng G, Zhang H, Tai M, et al. Assessment of hindlimb motor recovery after severe thoracic spinal cord injury in rats: classification of CatWalk XT gait analysis parameters. Neural Regen Res. 2023;18(5):1084–1089.
20. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–1182.
21. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40(8):913–929.
22. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189–1232.
23. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002;38(4):367–378.
24. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785–794.
25. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. LightGBM: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems. 2017;30.
26. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer; 2009.
27. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. 2017;30.
28. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56–67.
29. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2nd ed. 2022.
30. Hooker G, Mentch L, Zhou S. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. Stat Comput. 2021;31:82.
31. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010;26(10):1340–1347.
32. McCullagh P. Regression models for ordinal data. J R Stat Soc Series B. 1980;42(2):109–142.
33. Agresti A. Analysis of Ordinal Categorical Data. 2nd ed. Hoboken: Wiley; 2010.
34. Poon PC, Gupta D, Shoichet MS, Tator CH. Clip compression model is useful for thoracic spinal cord injuries: histologic and functional correlates. Spine (Phila Pa 1976). 2007;32(25):2853–2859.
35. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91.
36. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019:2623–2631.
37. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861–874.
38. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–138.
39. Field CA, Welsh AH. Bootstrapping clustered data. J R Stat Soc Series B Stat Methodol. 2007;69(3):369–390.
40. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. Biometrics. 1997;53(2):567–578.
41. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230.

42. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016;74:167–176.
43. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning. New York: Association for Computing Machinery; 2005:625–632.
44. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd ed. Cham: Springer; 2019.
45. Harrell FE Jr. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2nd ed. Cham: Springer; 2015.
46. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78(1):1–3.
47. Hastie T, Tibshirani R. Generalized additive models. Stat Sci. 1986;1(3):297–310.
48. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: Advances in Neural Information Processing Systems. 2018;31.
49. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–2830.
50. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
51. Moons KGM, Damen JAAG, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505.
52. Percie du Sert N, Hurst V, Ahluwalia A, Alam S, Avey MT, Baker M, et al. The ARRIVE guidelines 2.0: updated guidelines for reporting animal research. PLoS Biol. 2020;18(7):e3000410.
53. Percie du Sert N, Ahluwalia A, Alam S, Avey MT, Baker M, Browne WJ, et al. Reporting animal research: explanation and elaboration for the ARRIVE guidelines 2.0. PLoS Biol. 2020;18(7):e3000411.
54. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–58.
55. Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage. 2018;180:68–77.
56. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, TRIPOD: the TRIPOD statement. Ann Intern Med. 2015;162(1):55–63.