Supplementary methods

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

Analytical populations in the study workflow

Layer	Records	Animals	Analytical role
Complete source manifest	2,101	451	Five source experiments: SHH, VEGF, BMDM, IL4, and NK1.
Quality-control-retained population	1,524	387	Post-baseline source records meeting identifier and forepaw-stance availability rules.
Prespecified controls-only cohort	396	115	Comparative 26-feature modeling across XGBoost, LightGBM, and CatBoost with animal-grouped internal validation.
Three-feature analysis cohort	396	115	Post-selection nested-CV sensitivity analysis and recovery-index construction.

S1. Study scope, sources, and analytical target

S1.1 Study design and intended use

This supplementary document expands the methods for the post-acquisition machine-learning analysis reported in the main manuscript. We analyzed CatWalk XT recordings from five preclinical traumatic spinal cord injury (tSCI) experiments in Wistar rats. We performed no new intervention, allocation, or outcome acquisition.

We classified contemporaneous, time-defined recovery phase rather than predicting a future event or an independently adjudicated biological recovery state. We intended the models for internal research classification and feature analysis in control animals drawn from the pooled experimental protocol represented here.

S1.2 Source manifest and provenance

Workflow 00 read the Analysis_Ready worksheet for five source experiments and recorded file hashes, dimensions, and column inventories (Supplementary Table S1). The source manifest contained 2,101 rows from 451 animals across the five source experiments (SHH, VEGF, BMDM, IL4, and NK1). Workflow 01 mapped the source columns to a 110-column schema and added source_id, source_row_number, and globally unique row_id provenance fields.

S1.3 Ethical and source-method context

The source documentation identified animal-study approval G-211/15 under German animal-protection requirements. As described in the main manuscript, the source studies used a T9 clip contusion/compression injury, standardized perioperative care, seven days of CatWalk acclimation, and manual review of automatically classified paw prints. The computational analysis did not infer randomization, blinding, or sex composition when these items were not recorded in the source files.

S2. Paw aggregation, quality control, and cohort construction

S2.1 Bilateral paw compartments

Because the clip injury was intended to be laterally symmetrical, Workflow 02 combined right-front and left-front measurements into a forepaw (FP) compartment and right-hind and left-hind measurements into a hindpaw (HP) compartment. Under the prespecified missing_policy = available rule, we calculated the

arithmetic mean when both paws were observed, retained the single measurement when only one paw was observed, and returned a missing value when neither paw was observed. Run-average speed remained a global trial-level variable.

S2.2 Prespecified exclusions

We applied three prespecified exclusion rules: missing experiment, animal, or timepoint identifiers; baseline timepoint; and missing aggregated fp_stand_s (Supplementary Table S2). Of the 2,101 source rows, we excluded 534 baseline rows and 43 rows with the required measurement missing. The remaining 1,524 rows were retained in the row-level quality-control audit.

S2.3 Group and phase harmonization

Workflow 03 derived group_type from version-controlled, experiment-specific treatment mappings. We assigned recovery phase using the corresponding experiment-specific timepoint maps (Supplementary Table S3): available day-3 and day-7 observations were assigned to Early; days 14 and 21 were assigned to Middle where available; and days 28 and 35 were assigned to Late where available. Baseline day 0 had no ordinal phase.

S2.4 Frozen controls-only cohort

Workflow 04 retained rows from one of the five source experiments when group_type = control, timepoint > 0, the phase was Early, Middle, or Late, factorized_animal was nonmissing, and the row had passed upstream quality control. The frozen controls-only cohort contained 396 records from 115 animals: 175 Early records from 93 animals, 116 Middle records from 58 animals, and 105 Late records from 33 animals. Because an animal could contribute records to more than one phase, the phase-specific animal counts are not additive.

S3. Feature engineering

S3.1 Raw model inputs

The modeling pipeline received 19 paw-aggregated raw inputs: run-average speed and, for both FP and HP, intensity, maximum intensity, 15-pixel intensity, stance, swing, print area, maximum contact area, stride length, and duty cycle (Supplementary Table S4). We excluded identifiers, experiment, timepoint, group, recovery phase, and quality-control fields from the predictor matrix.

S3.2 Fold-local 26-feature transformer

Within each training partition, the predefined transformer generated 26 engineered variables. Peak-loading indices used the mean and sample standard deviation estimated from that training partition. Ratio calculations divided by denominator + epsilon, where epsilon = 1 x 10^-8, and infinite results were replaced with missing values. Supplementary Table S5 gives the complete feature dictionary.

Feature transformation was the first step in every scikit-learn pipeline. During inner validation, we fitted the transformer only on the inner-training animals. For final fitting within an outer fold, we refitted it on all outer-training animals and applied it unchanged to the corresponding outer-test animals.

S4. Missing data

Missingness primarily reflected trials in which a hindpaw could not be detected because hindlimb motor function was absent or severely impaired. The 26-feature matrix contained 4,179 missing values among 10,296 cells (40.6%), and 113 of 396 records were complete across all 26 features. Missingness was 56.6% in Early, 32.7% in Middle, and 22.6% in Late. The three-feature matrix contained 452 missing values among 1,188 cells (38.0%), and 162 records were complete across all three features (Supplementary Table S6).

We did not apply explicit imputation before fitting XGBoost, LightGBM, or CatBoost; missing engineered values were passed to each classifier's native missing-value mechanism. The GAM surrogate used median imputation fitted separately within each training partition.

S5. Comparative ordinal modeling and nested validation

S5.1 Ordinal decomposition

For the ordered outcome Early < Middle < Late, we fitted two binary estimators: q1 = P(Y > Early) and q2 = P(Y > Middle) = P(Late). We enforced ordinal consistency by setting q2 = min(q2, q1), then reconstructed P(Early) = 1 - q1, P(Middle) = q1 - q2, and P(Late) = q2. For hard classification, we assigned each record to the phase with the largest reconstructed probability.

S5.2 Outer and inner folds

Workflow 05 generated five outer folds by stratifying on recovery phase, grouping on factorized_animal, and balancing experiment-by-phase composition where feasible (Supplementary Table S7). The outer-test folds contained 79, 79, 78, 80, and 80 records. Within every outer-training partition, we performed five-fold grouped inner validation. The same frozen outer assignments were used for every framework and for the later three-feature analysis.

S5.3 Hyperparameter optimization

For each framework in each outer fold, we ran an independent 100-trial Optuna study, giving 15 studies in total (Supplementary Table S8). The optimization objective was the mean cumulative area under the receiver operating characteristic curve (AUROC), calculated as the arithmetic mean of the Early and Late AUROCs across five grouped inner folds. We selected parameters using outer-training data only, fitted a fresh pipeline to the complete outer-training partition, and evaluated it once on the untouched outer-test animals.

S5.4 Classifier comparison and equal-weight ensemble refits

The XGBoost, LightGBM, and CatBoost pipelines each used all 26 engineered features for the three-framework comparison. In paired out-of-fold mean-AUROC comparisons, only the CatBoost–XGBoost interval excluded zero; the other two intervals included zero. This comparison did not select a single-framework application artifact. After Workflow 12 defined the locked three-feature panel and Workflow 13 specified equal one-third weights with Platt calibration for all components, Workflow 15 tuned and refitted all three three-feature frameworks on the complete 396-record cohort and exported their equal-weight ensemble for application. These full-data refits were not used to estimate performance.

S6. Performance, uncertainty, calibration, and sensitivity

S6.1 Metrics and clustered uncertainty

The Early AUROC distinguished Early from Middle/Late, whereas the Late AUROC distinguished Early/Middle from Late; the mean cumulative AUROC was their arithmetic mean. Pooled out-of-fold estimates were calculated by concatenating the five mutually exclusive outer-test sets. We also reported accuracy, balanced accuracy, ordinal mean absolute error, log loss, and cumulative Brier scores (Supplementary Table S9).

Workflow 06 generated 5,000 bootstrap replicates by resampling factorized_animal clusters within experiment. We applied the same sampled animals to every framework so that pairwise comparisons remained paired (Supplementary Table S10). Two-sided sign-tail bootstrap p values were descriptive and were not adjusted for multiple comparisons.

S6.2 Cross-fitted calibration

Workflow 07 evaluated no calibration, logistic (Platt) calibration, and isotonic calibration separately for q1 and q2. We fitted calibrators to grouped inner out-of-fold predictions and applied them to the untouched outer-test fold. Under the conservative selection rule, calibration was selected only when the clustered 95% confidence interval for improvement in mean cumulative Brier score excluded zero and pooled log loss did not worsen. Logistic calibration met these criteria for XGBoost and LightGBM; no calibration was selected for CatBoost (Supplementary Table S11).

S6.3 Prediction sensitivity

Workflow 08 compared uncalibrated predictions and framework-selected calibration under row-weighted pooled, equal-animal-weighted, and equal-outer-fold estimands, with animal-clustered uncertainty. We did not evaluate alternative quality-control rules, paw-missingness definitions, or phase mappings because each alternative would require a separately frozen cohort and a complete nested-model rerun. The validation estimand was performance in new animals drawn from the represented pooled protocol. Experiment was used for balancing rather than as a held-out validation unit because phase support differed structurally across experiments.

S7. Native TreeSHAP, feature ablation, and consensus

Workflow 09 computed native TreeSHAP contributions using the XGBoost, LightGBM, and CatBoost tree-library interfaces. We explained only held-out outer-fold records and verified additivity on the raw-margin scale for both cumulative thresholds.

Workflow 10 treated all 26 engineered features as the ablation universe. We removed each feature in turn, refitted the corresponding pipeline on the same outer-training data using the saved base-estimator configuration, and evaluated it on the same outer-test records. Hyperparameters were not reoptimized for individual feature removals.

Within each of 15 framework-by-outer-fold analyses, Workflow 11 ranked features separately by held-out TreeSHAP importance and by held-out mean-cumulative-AUROC loss after ablation (Supplementary Tables S13 and S14). The primary combined ranking pooled the resulting 30 ranks 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 SHAP-heavy or ablation-heavy weightings were descriptive and did not determine selection. Agreement was stronger among SHAP judges (Kendall W = 0.686) than among ablation judges (W = 0.168); SHAP-consensus versus ablation-consensus concordance was modest (Spearman rho = 0.217; Kendall tau-b = 0.169). Forepaw intensity density was the clearest shared signal: it ranked first and produced an average mean cumulative AUROC loss of 0.0224 when removed.

S8. Post-selection three-feature nested cross-validation analysis

The combined ranking selected fp_intensity_density, fp_intensity_loading_fraction, and hp_loading_rate. Workflow 12 fixed this three-feature panel and ran a fresh five-by-five grouped nested cross-validation analysis with 100 trials for every framework-by-outer-fold combination (Supplementary Table S15).

Because we selected the feature panel using out-of-fold evidence from the complete 115-animal controls-only cohort, the three-feature rerun is a post-selection sensitivity analysis rather than independent external or confirmatory validation.

S9. Ensemble recovery index

The three-feature calibration analysis selected logistic (Platt) calibration for CatBoost; no calibration was selected for LightGBM or XGBoost under the conservative probability-accuracy rule. For the separate purpose

of constructing comparable recovery-index components, however, Workflow 13 applied Platt calibration to all three three-feature frameworks. For each framework, the recovery index (RI) was defined as RI = 50 × P(Middle) + 100 × P(Late). The ensemble recovery index was the equal-weight mean of the CatBoost, LightGBM, and XGBoost indices (Supplementary Table S16).

Every index value was calculated from an outer-test prediction. Calibrators were fitted to inner out-of-fold predictions without using the corresponding outer-test animal. The resulting 0-to-100 score represents expected ordinal class position, not percentage biological recovery.

S10. Generalized additive model (GAM) recovery-index surrogate

Workflow 14 fitted a penalized additive-spline regression surrogate to the ensemble recovery index rather than to the observed Early, Middle, or Late labels. The surrogate used the same three engineered predictors. Median imputation, cubic-spline construction, scaling, and penalized fitting were all restricted to the active training partition. Candidate knot counts were 4, 5, 6, and 7, and candidate alpha values were 0.01, 0.1, 1, 10, and 100.

We evaluated the surrogate only for fidelity to the out-of-fold ensemble index (Supplementary Table S17). An all-data surrogate was fitted for application, whereas all reported fidelity estimates remained out of fold.

S11. Classifier-ensemble artifact and interpretation scope

Workflow 15 exported an equal-weight ensemble artifact comprising Platt-calibrated three-feature XGBoost, LightGBM, and CatBoost components refitted to the complete 396-record cohort for application to new control-animal records measured under the pooled experimental protocol represented in this study. These full-data application components were not used for performance estimation and have not been validated for unseen laboratories, treatment effects, or biological recovery endpoints.

The three-feature equal-weight ensemble recovery index is the second analytical stage, and its GAM surrogate approximates the ensemble output. Classifier-comparison artifacts, ensemble-index artifacts, and surrogate artifacts remain separately identified in tables, figures, and repository directories; no classifier framework is ranked or selected.

S12. Software and reproducibility

We ran the canonical workflow in Python 3.10.8 using scikit-learn 1.7.2, XGBoost 3.2.0, LightGBM 4.7.0, CatBoost 1.2.10, Optuna 4.9.0, pandas 2.3.3, NumPy 2.2.6, and joblib 1.5.3. SciPy 1.15.3 and Matplotlib 3.10.9 were locked dependencies. The base random seed was 0 (Supplementary Table S18). Reporting and risk-of-bias considerations were informed by ARRIVE 2.0, TRIPOD+AI, and PROBAST+AI [50–53].

Each workflow from 00 through 15 writes a completion manifest containing its configuration, lineage, and file fingerprints (Supplementary Table S19). Portable manifest exports remove machine-specific absolute paths. The classifier-ensemble bundle includes all three co-equal components, a validated loader, and per-component plus ensemble SHA-256 verification.

S13. Limitations and data availability

The main limitations are internal validation only; a timepoint-derived target; stage-dependent missingness; repeated observations; hyperparameter tuning and feature-panel derivation within one controls-only cohort; lack of control-cohort sex metadata; and no independent laboratory or biological validation. We did not use experiment-blocked validation because the source experiments did not provide comparable phase support. Consequently, transport to other laboratories, strains, sexes, injury severities, treatment groups, acquisition systems, or processing pipelines remains unknown.

The analysis code and aggregate analytical artifacts are available through the curated repository at https://gitlab.com/sergienko.michael/sci. Raw CatWalk workbooks and derived animal-level tables are not publicly distributed. Requests should be directed to the corresponding author and remain subject to institutional approval, applicable ethics requirements, and any required data-use agreement.