Supplementary tables

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

Supplementary Table S1. Source-experiment manifest

Experiment	Status	Rows	Columns
SHH	Included	916	136
VEGF	Included	388	134
BMDM	Included	221	332
IL4	Included	393	278
NK1	Included	183	189
Total	5 experiments	2,101	-

Supplementary Table S2. Quality control and cohort derivation

Cohort step	Records	Animals	Definition
Canonical source dataset	2,101	451	Five source experiments after Workflow 01 canonicalization
Baseline excluded	534	-	Timepoint = 0; no ordinal phase
Missing measurement excluded	43	-	Missing fp_stand_s
Workflow 02 QC-retained	1,524	387	Post-baseline source rows passing QC
Controls-only cohort	396	115	QC-retained controls with an assigned phase
QC-retained non-control rows	1,128	-	840 experimental; 281 sham; 7 unclassified

Supplementary Table S3. Controls-only phase and timepoint composition

Phase	Post-injury day	Records	Animals
Early	Day 3	68	68
Early	Day 7	107	74
Middle	Day 14	59	45
Middle	Day 21	57	30
Late	Day 28	64	33
Late	Day 35	41	15

Supplementary Table S4. Raw predictor inputs

Input feature	Scope

Input feature	Scope
run_average_speed_cm_per_s	Global
fp_intensity	Forepaw
hp_intensity	Hindpaw
fp_max_intensity	Forepaw
hp_max_intensity	Hindpaw
fp_intensity_of_the_15_most_intense_pixels	Forepaw
hp_intensity_of_the_15_most_intense_pixels	Hindpaw
fp_stand_s	Forepaw
hp_stand_s	Hindpaw
fp_swing_s	Forepaw
hp_swing_s	Hindpaw
fp_print_area_cm2	Forepaw
hp_print_area_cm2	Hindpaw
fp_max_contact_area_cm2	Forepaw
hp_max_contact_area_cm2	Hindpaw
fp_stride_length_cm	Forepaw
hp_stride_length_cm	Hindpaw
fp_duty_cycle_pct	Forepaw
hp_duty_cycle_pct	Hindpaw

Supplementary Table S5. Engineered-feature dictionary

Engineered feature	Formula
hp_peak_loading_index	(z(HP maximum intensity) + z(HP 15-pixel intensity)) / 2
fp_peak_loading_index	(z(FP maximum intensity) + z(FP 15-pixel intensity)) / 2
fp_intensity_loading_fraction	FP intensity / (FP intensity + HP intensity + epsilon)
hp_intensity_loading_fraction	HP intensity / (FP intensity + HP intensity + epsilon)
fp_maxintensity_loading_fraction	FP maximum intensity / (FP maximum intensity + HP maximum intensity + epsilon)
hp_maxintensity_loading_fraction	HP maximum intensity / (FP maximum intensity + HP maximum intensity + epsilon)
hp_loading_speed	HP intensity / (run-average speed + epsilon)
fp_loading_speed	FP intensity / (run-average speed + epsilon)
hp_functional_output	HP intensity x HP print area x HP stance
fp_functional_output	FP intensity x FP print area x FP stance

Engineered feature	Formula
fp_intensity_density	FP intensity / (FP maximum contact area + epsilon)
hp_intensity_density	HP intensity / (HP maximum contact area + epsilon)
fp_integrated_loading	FP intensity x FP stance
hp_integrated_loading	HP intensity x HP stance
fp_loading_rate	FP intensity / (FP stance + epsilon)
hp_loading_rate	HP intensity / (HP stance + epsilon)
fp_propulsion_time	FP stride length / (FP stance + epsilon)
hp_propulsion_time	HP stride length / (HP stance + epsilon)
stand_symmetry	|FP stance - HP stance|
swing_symmetry	|FP swing - HP swing|
duty_cycle_symmetry	|FP duty cycle - HP duty cycle|
fp_temporal_contact	FP print area x FP stance
hp_temporal_contact	HP print area x HP stance
hp_propulsion_loading	HP stride length / (HP intensity + epsilon)
hp_stride_length_per_cycle	HP stride length / (HP stance + HP swing + epsilon)
fp_stride_length_per_cycle	FP stride length / (FP stance + FP swing + epsilon)

For ratios, epsilon = 1 × 10^-8; peak-loading z statistics are estimated in the active training partition; infinite outputs are set to missing.

Supplementary Table S6. Engineered-feature missingness

Subset	Cells	Missing	Missing, %	Complete records, n/N
All records (26 features)	10,296	4,179	40.6	113/396
Early (26 features)	4,550	2,577	56.6	-
Middle (26 features)	3,016	985	32.7	-
Late (26 features)	2,730	617	22.6	-
All records (3 features)	1,188	452	38.0	162/396
Early (3 features)	525	294	56.0	28/175
Middle (3 features)	348	96	27.6	60/116
Late (3 features)	315	62	19.7	74/105

Supplementary Table S7. Outer-fold composition

Fold	Records	Animals	Early	Middle	Late
1	79	22	34	23	22
2	79	23	36	23	20

Fold	Records	Animals	Early	Middle	Late
3	78	23	34	23	21
4	80	23	34	24	22
5	80	24	37	23	20

Each animal appears in one outer-test fold; assignments were fixed across frameworks and reused for the three-feature rerun.

Supplementary Table S8. Optuna search spaces

Framework	Parameter(s)	Range
XGBoost	n_estimators; max_depth	50-400; 2-10
	learning_rate	0.001-0.3 (log)
	subsample; colsample_bytree	0.5-1.0; 0.5-1.0
	reg_alpha; reg_lambda; gamma	0-5; 0-5; 0-5
	min_child_weight	1-10
LightGBM	n_estimators; max_depth	50-500; 2-10
	learning_rate	0.001-0.3 (log)
	num_leaves exponent	2-7
	subsample; colsample_bytree	0.5-1.0; 0.5-1.0
	reg_alpha; reg_lambda	0-5; 0-10
	min_child_samples	5-50
	min_child_weight	0.001-10 (log)
CatBoost	iterations; depth	50-500; 2-10
	learning_rate	0.001-0.3 (log)
	bagging_temperature	0-10
	border_count	32-255
	l2_leaf_reg	1-20
	random_strength	0-10

Each of 15 studies used 100 trials; the objective was mean cumulative AUROC across five animal-grouped inner folds.

Supplementary Table S9. Out-of-fold performance across the three 26-feature classifiers

Framework	Metric	Estimate	95% CI
CatBoost	Early AUROC	0.8706	[0.8306, 0.9084]
	Late AUROC	0.8516	[0.8178, 0.8870]
	Mean cumulative AUROC	0.8611	[0.8324, 0.8893]
CatBoost (continued)	Accuracy	0.6439	[0.5938, 0.6957]

Framework	Metric	Estimate	95% CI
	Balanced accuracy	0.6056	[0.5495, 0.6609]
	Ordinal MAE	0.4066	[0.3444, 0.4679]
	Mean cumulative Brier score	0.1405	[0.1245, 0.1565]
LightGBM	Early AUROC	0.8594	[0.8120, 0.9034]
	Late AUROC	0.8304	[0.7926, 0.8667]
	Mean cumulative AUROC	0.8449	[0.8106, 0.8775]
	Accuracy	0.6364	[0.5848, 0.6872]
	Balanced accuracy	0.5962	[0.5398, 0.6505]
	Ordinal MAE	0.4571	[0.3879, 0.5319]
	Mean cumulative Brier score	0.1486	[0.1347, 0.1629]
XGBoost	Early AUROC	0.8519	[0.8024, 0.8976]
	Late AUROC	0.8174	[0.7818, 0.8537]
	Mean cumulative AUROC	0.8347	[0.8009, 0.8687]
	Accuracy	0.6212	[0.5696, 0.6739]
	Balanced accuracy	0.5803	[0.5241, 0.6366]
	Ordinal MAE	0.4672	[0.3889, 0.5448]
	Mean cumulative Brier score	0.1544	[0.1387, 0.1702]

All 396 records are mutually exclusive outer-test observations; 95% confidence intervals (CIs) use 5,000 experiment-stratified, animal-clustered bootstrap replicates.

Supplementary Table S10. Paired differences in mean cumulative AUROC

Framework A	Framework B	A - B	95% CI	p value
CatBoost	LightGBM	0.0162	[-0.0021, 0.0346]	0.0784
CatBoost	XGBoost	0.0264	[0.0086, 0.0447]	0.0040
LightGBM	XGBoost	0.0102	[-0.0062, 0.0258]	0.2228

Paired bootstrap was experiment stratified and animal clustered; two-sided sign-tail p values are descriptive and unadjusted.

Supplementary Table S11. Cross-fitted calibration selection

Framework	Selected method	Δ mean Brier score [95% CI]
CatBoost	No calibration	0.0012 [-0.0015, 0.0040]
LightGBM	Logistic (Platt)	0.0065 [0.0007, 0.0126]
XGBoost	Logistic (Platt)	0.0081 [0.0022, 0.0147]

Positive values indicate lower mean cumulative Brier score after logistic calibration. CatBoost remained uncalibrated because its CI included zero and log loss worsened.

Supplementary Table S12. Out-of-fold confusion matrices for the three 26-feature classifiers

Framework	Observed	Predicted	Count	Row, %
CatBoost	Early	Early	149	85.1
		Middle	20	11.4
		Late	6	3.4
	Middle	Early	40	34.5
		Middle	49	42.2
		Late	27	23.3
	Late	Early	14	13.3
		Middle	34	32.4
		Late	57	54.3
LightGBM	Early	Early	151	86.3
		Middle	15	8.6
		Late	9	5.1
	Middle	Early	46	39.7
		Middle	40	34.5
		Late	30	25.9
	Late	Early	28	26.7
		Middle	16	15.2
		Late	61	58.1
XGBoost	Early	Early	149	85.1
		Middle	17	9.7
		Late	9	5.1
	Middle	Early	44	37.9
		Middle	38	32.8
		Late	34	29.3
	Late	Early	26	24.8
		Middle	20	19.0
		Late	59	56.2

Supplementary Table S13. Combined SHAP-ablation feature ranking

Overall rank	Feature	SHAP rank	Mean SHAP rank	Δ mean cumulative AUROC (baseline - ablated)
1	fp_intensity_density	1	1.00	0.02243
2	fp_intensity_loading_fraction	2	4.20	0.00159
3	hp_loading_rate	5	6.60	0.00349
4	fp_loading_speed	3	5.80	0.00063
5	hp_loading_speed	8	8.60	0.00098
6	hp_temporal_contact	9	8.93	0.00300
7	hp_integrated_loading	10	9.13	0.00138
8	fp_temporal_contact	4	6.33	0.00170
9	fp_peak_loading_index	7	7.87	-0.00106
10	fp_loading_rate	11	11.40	0.00076
11	fp_functional_output	6	7.87	-0.00134
12	hp_intensity_density	12	14.33	0.00173
13	fp_propulsion_time	14	15.00	0.00185
14	hp_propulsion_loading	16	15.77	0.00248
15	hp_intensity_loading_fraction	13	14.53	0.00093
16	fp_stride_length_per_cycle	17	17.07	0.00146
17	fp_integrated_loading	19	17.67	0.00073
18	hp_functional_output	15	15.23	-0.00144
19	fp_maxintensity_loading_fraction	18	17.13	0.00123
20	stand_symmetry	23	21.60	0.00137
21	duty_cycle_symmetry	24	22.20	0.00183
22	swing_symmetry	22	19.53	0.00075
23	hp_stride_length_per_cycle	26	23.37	0.00259
24	hp_maxintensity_loading_fraction	20	18.47	-0.00070
25	hp_peak_loading_index	21	18.70	-0.00027
26	hp_propulsion_time	25	22.67	0.00161

Supplementary Table S14. Feature-ranking agreement across 26 features

Comparison	Statistic	Value
SHAP across 15 fold-model judges	Kendall W	0.686

Comparison	Statistic	Value
Ablation across 15 fold-model judges	Kendall W	0.168
All 30 SHAP and ablation judges	Kendall W	0.278
Method-consensus rankings	Kendall W	0.701
SHAP vs ablation consensus	Spearman rho	0.217
	Kendall tau-b	0.169
	Rank R-squared	0.047

Supplementary Table S15. Post-selection three-feature panel and nested cross-validation performance

Section	Rank / framework	Feature / scope	Early AUROC	Late AUROC	Mean cumulative AUROC [95% CI]
Panel	1	fp_intensity_density	-	-	-
Panel	2	fp_intensity_loading_fraction	-	-	-
Panel	3	hp_loading_rate	-	-	-
Performance	CatBoost	All three features	0.880	0.833	0.856 [0.826-0.886]
Performance	LightGBM	All three features	0.874	0.837	0.856 [0.826-0.886]
Performance	XGBoost	All three features	0.878	0.840	0.859 [0.830-0.888]

Fresh 5 × 5 animal-grouped nested cross-validation used all three features. CIs are animal clustered. Because selection used the same 115 animals, results are post-selection sensitivity estimates, not independent validation.

Supplementary Table S16. Recovery index by phase and framework

Index source	Phase	Mean	SD	Median
CatBoost	**Early**	22.77	16.81	16.87
LightGBM		24.59	16.11	18.33
XGBoost		25.06	15.31	17.58
Ensemble		24.14	15.92	17.01
CatBoost	**Middle**	49.66	23.02	50.19
LightGBM		49.32	21.87	46.82
XGBoost		48.84	21.06	45.35
Ensemble		49.28	21.81	47.18
CatBoost	**Late**	64.58	21.52	68.50
LightGBM		63.78	19.78	69.24
XGBoost		63.21	19.15	68.65

Index source	Phase	Mean	SD	Median
Ensemble		63.86	19.94	70.40

Phase sizes (records/animals): Early 175/93; Middle 116/58; Late 105/33. Ensemble means (95% CI): Early 24.1 (21.9-26.7); Middle 49.3 (44.1-54.4); Late 63.9 (58.8-68.8). Equal-weight Platt-calibrated ensemble: Early, Late, and mean cumulative AUROCs 0.865, 0.828, and 0.846; Spearman rho with phase = 0.655.

Supplementary Table S17. Generalized additive model (GAM) surrogate fidelity

Metric	Out-of-fold estimate
R-squared	0.9566
RMSE	5.260
MAE	4.007
Mean error	-0.348
Max absolute error	16.456
Within 5 index points	71.0%
Within 10 index points	91.9%
Pearson r	0.9783
Spearman rho	0.9688
Concordance correlation coefficient	0.9782
Agreement slope	0.9844

Target: Workflow 13 ensemble index. Values measure surrogate fidelity, not independent recovery-phase validation.

Supplementary Table S18. Software environment and classifier-ensemble artifact

Component	Version / value	Role
Python	3.10.8	Workflows
scikit-learn	1.7.2	Pipelines; grouped CV; metrics
XGBoost	3.2.0	Three-classifier comparison; co-equal ensemble component
LightGBM	4.7.0	Three-classifier comparison; co-equal ensemble component
CatBoost	1.2.10	Three-classifier comparison; co-equal ensemble component
Optuna	4.9.0	100-trial search
pandas	2.3.3	Data processing
NumPy	2.2.6	Numerics
joblib	1.5.3	Persistence; parallelism
SciPy	1.15.3	Bootstrap; correlations
Matplotlib	3.10.9	Figures
Base seed	0	Folds; derived seeds

Component	Version / value	Role
Ensemble frameworks	XGBoost; LightGBM; CatBoost	Equal one-third weights; no framework selection
Features per component	3	Locked Workflow-12 panel
Component calibration	Platt/logistic for all	Workflow-13 comparability policy
Artifact SHA-256	Recorded per component and ensemble	Workflow-15 run manifest

Supplementary Table S19. Workflow outputs

Step	Purpose	Output
00-04	Cohort preparation	Frozen 396-record cohort; QC audits
05	Three-classifier 26-feature nested comparison	15 fold models; out-of-fold predictions
06	Clustered inference	Metrics, CIs, paired comparisons
07	Cross-fitted calibration	Selected method per framework
08	Prediction sensitivity	Estimand comparisons
09	Native TreeSHAP	Held-out attributions; rankings
10-11	Ablation and consensus	26-feature effects; consensus ranks
12	Three-feature nested cross-validation	Fresh three-feature out-of-fold predictions
13	Recovery index	Framework indices; equal-weight ensemble
14	GAM surrogate	Out-of-fold fidelity; deployable spline model
15	Classifier-ensemble fitting and export	Three Platt-calibrated three-feature components; equal-weight ensemble artifact

Experiment balanced the folds but was not a held-out validation unit because phase support differed across studies.