Supplementary checklists

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

Scope. ARRIVE 2.0 applies to the source animal studies and secondary analysis. TRIPOD+AI and PROBAST+AI are adapted reporting/appraisal frameworks only; no formal compliance or validated PROBAST+AI rating is claimed. Unavailable source-study details are identified rather than inferred.

Status key. Complete/Confirmed/Yes = reported; Partly complete/Some concern/Qualified = limited but discussed; other labels state the applicable restriction. Statuses describe reporting, not external verification.

Checklist S1. Adapted TRIPOD+AI reporting map

TRIPOD+AI item	Requirement	Status	Location / response
1. Title	Model study, population, and outcome.	**Complete**	Title identifies interpretable machine learning, gait after spinal cord injury, and time-defined recovery-phase classification.
2. Abstract	Objectives, data, methods, results, and conclusions.	**Partly complete**	Abstract reports the cohort, predictors, three-classifier comparison, ensemble-index derivation, surrogate analysis, and validation limits.
3a. Background	Context and rationale.	**Complete**	Introduction: rationale for multivariable, interpretable ordinal gait modeling.
3b. Target population and intended purpose	Target population, purpose, and users.	**Partly complete**	Main manuscript; Supplementary Methods S1: control Wistar rats under the represented CatWalk protocol; intended users are not stated explicitly.
3c. Health inequalities / subgroup concerns	Relevant subgroup or inequality concerns.	**Adapted / limited**	Human inequalities are not applicable. Sex-specific assessment was unavailable; transportability across sex, strain, experiment, injury context, and missingness is discussed.
4. Objectives	Development/evaluation objective.	**Complete**	Abstract/Introduction: development, internal validation, and comparison of three ordinal tree classifiers, followed by feature interpretation, a three-feature classifier rerun, equal-weight ensemble recovery-index construction, and GAM-surrogate analysis.
5a. Data source	Data source and representativeness.	**Complete**	Methods 3.1; Supplementary Methods S1; Tables S1-S3: 2,101 source rows/451 animals from five source experiments, and a frozen 396-record/115-animal cohort.
5b. Dates of collected data	Collection dates and follow-up.	**Partly complete**	Calendar dates were unavailable; baseline and days 1, 3, 7, 14, 21, 28, and 35 are reported where present, with experiment-specific phase maps.
6a. Study setting	Setting, centres, and location.	**Partly complete**	Preclinical CatWalk SCI study in Germany; laboratory/centre and calendar period are not stated.
6b. Eligibility criteria	Eligibility criteria.	**Complete**	Methods 3.2; Supplementary Methods S2; Table S2: post-baseline, QC-retained control records with valid animal ID and assigned phase.
6c. Treatments received	Treatments and analytical handling.	**Complete / adapted**	Controls only; treatment labels were excluded from predictors.
7. Data preparation	Preprocessing, QC, and harmonization.	**Complete**	Methods 3.1-3.4; Supplementary Methods S1-S4; Tables S1-S6: provenance, bilateral aggregation, QC, harmonization, cohort freezing, and fold-local features.
8a. Outcome definition	Outcome definition, timing, and rationale.	**Complete**	Methods 3.2; Supplementary Methods S2.3: Early = IL4 day 1, VEGF day 3, and all available day 7 observations; Middle = days 14/21; Late = days 28/35 where present. The outcome is contemporaneous, not prognostic or biological.
8b. Outcome assessor qualifications	Outcome-assessor qualifications.	**Not applicable**	Deterministic metadata-derived outcome.
8c. Blinding of outcome assessment	Outcome-assessment blinding.	**Not applicable**	No subjective assessment; timepoint and recovery_phase defined labels but were excluded from predictors.
9a. Predictor selection	Initial predictor selection.	**Complete**	Methods 3.3/3.5; Supplementary Methods S3; Tables S4-S5: 19 raw inputs to 26 prespecified features; 26-feature classifier comparison followed by a post-selection three-feature ensemble-index analysis.
9b. Predictor definitions	Predictor definitions and transformations.	**Complete**	Supplementary Methods S3; Tables S4-S5: inputs, units, bilateral aggregation, formulas, epsilon, and fold-local z statistics.
9c. Predictor assessor qualifications	Predictor-assessor qualifications.	**Complete / limited**	CatWalk outputs were automatically classified, then corrected and validated by three researchers; masking and inter-rater reliability were unavailable.
10. Sample size	Sample-size basis.	**Partly complete**	396 records/115 animals; no prospective calculation; all eligible records used; animal-clustered uncertainty.
11. Missing data	Missing-data handling.	**Complete**	Methods 3.4; Supplementary Methods S4; Table S6: 4,179/10,296 cells missing (40.6%); trees used native handling, and the GAM used fold-local median imputation.
12a. Data use / partitioning	Development/evaluation partitioning.	**Complete**	Methods 3.5; Supplementary Methods S5.2; Table S7: five animal-grouped outer folds and five grouped inner folds; experiment-phase balancing where feasible.

TRIPOD+AI item	Requirement	Status	Location / response
12b. Predictor handling	Transformations, rescaling, and functional form.	**Complete**	Fold-local engineering/statistics; no global scaling or imputation; GAM preprocessing was fold-local.
12c. Model type, tuning, internal validation	Model type, tuning, and internal validation.	**Complete**	Methods 3.5; Supplementary Methods S5; Table S8: ordinal CatBoost, LightGBM, and XGBoost; 100 Optuna trials per framework-fold; grouped nested cross-validation.
12d. Heterogeneity / clustering	Clustering and heterogeneity.	**Complete**	Animal-grouped splitting/inference; experiment used for balancing and summaries, not prediction. Estimand: new animals under the pooled protocol.
12e. Performance measures	Discrimination, calibration, error, and comparison measures.	**Complete**	Methods 3.6; Supplementary Methods S6-S10; Tables S9-S17: cumulative AUROCs, accuracy, balanced accuracy, ordinal MAE, Brier score, log loss, calibration, clustered CIs, and GAM fidelity.
12f. Model updating	Model updating after evaluation.	**Not applicable**	No updating; full-data refits of all three three-feature ensemble components were excluded from performance estimation.
12g. Prediction calculation	Prediction calculation and implementation.	**Complete**	Cumulative probabilities reconstruct Early/Middle/Late probabilities; RI = 50 × P(Middle) + 100 × P(Late). Pipelines and code: https://gitlab.com/sergienko.michael/sci
13. Class imbalance	Class-imbalance handling.	**Complete**	Early/Middle/Late: 175/116/105 records. No outcome resampling; grouped folds balanced phase where feasible.
14. Fairness	Fairness and subgroup assessment.	**Adapted / limited**	Human fairness is not applicable; sex was unavailable. Transport across sex, strain, severity, treatment, laboratory, and acquisition requires validation.
15. Model output	Outputs and thresholds.	**Complete**	Class probabilities and predicted phase from each classifier; equal-weight 0-100 ensemble recovery index. No clinical threshold.
16. Training vs evaluation differences	Development-versus-evaluation differences.	**Complete (internal)**	Animal-separated folds came from the same pooled protocol; estimates target new animals within that protocol, not external settings.
17. Ethical approval	Ethics approval.	**Complete (qualified)**	Ethics statement: source documentation identifies G-211/15 under German requirements; no unverified details were inferred.
18a. Funding	Funding and funder role.	**Complete**	Declarations: no external funding.
18b. Conflicts of interest	Conflicts of interest.	**Complete**	Declarations: none.
18c. Protocol	Protocol availability.	**Partly complete**	Methods, configurations, and versioned code are available; no separate prospective ML protocol.
18d. Registration	Registration.	**Not registered**	No prospective prediction-model registration reported.
18e. Data sharing	Data sharing.	**Complete (restricted)**	Data Availability; Supplementary Methods S13: aggregate outputs available; raw and animal-level data controlled.
18f. Code sharing	Code sharing.	**Complete**	https://gitlab.com/sergienko.michael/sci
19. Patient / public involvement	Patient/public involvement.	**Not applicable**	Preclinical secondary analysis; none.
20a. Participant / record flow	Animal/record flow.	**Complete**	Table S2: 2,101 source rows/451 animals; 1,524 QC-retained rows/387 animals; frozen cohort 396 records/115 animals; exclusions reported.
20b. Characteristics	Characteristics and missingness.	**Complete / limited**	Results 4.1; Tables S2-S6: cohort and missingness distributions. Wistar rats aged approximately 6–8 weeks and weighing 230–265 g; sex, individual age/weight, and calendar dates unavailable.
20c. Evaluation vs development distributions	Development/evaluation distributions.	**Partly complete (internal)**	Table S7 reports fold-level records, animals, and phase counts; predictor distributions by fold are not tabulated.
21. Model-development sample size per analysis	Sample size per analysis.	**Complete**	Tables S2–S3 and S7, together with the frozen fold-assignment artifacts.

TRIPOD+AI item	Requirement	Status	Location / response
22. Full model specification	Full model specification.	**Complete**	Tables S5, S8, S18-S19 and repository artifacts: formulas, search spaces, saved classifier components, ensemble metadata, dependencies, provenance, and fold parameters.
23a. Model performance	Performance estimates and uncertainty.	**Complete**	Results 4.2; main Table 3; Supplementary Table S9; Figures S1-S3. CatBoost mean cumulative AUROC 0.861 (95% CI 0.832–0.889); threshold and paired results reported.
23b. Cluster heterogeneity	Cluster heterogeneity.	**Partly complete**	Fold metrics and animal-clustered CIs are reported; experiment summaries are contextual because phase support differs.
24. Model updating	Model updating.	**Not applicable**	No updating performed.
25. Interpretation	Interpretation.	**Complete**	Discussion: phase classification is not biological recovery; attribution is not causality; the classifier comparison, three-feature sensitivity analysis, ensemble index, and surrogate are explicitly distinguished.
26. Limitations	Limitations and generalizability.	**Complete**	Discussion: internal validation, clustered sample, phase-dependent missingness, time-derived outcome, bilateral aggregation, post-selection analyses, and transportability.
27a. Poor-quality / unavailable input data	Handling poor/unavailable inputs.	**Complete**	Pipeline validates schema and transformations; tree models handle missingness natively; new schemas require revalidation.
27b. User interaction / expertise	Required user expertise.	**Partly complete**	Use requires expertise in CatWalk exports, preclinical SCI gait analysis, and time-defined—not prognostic—interpretation; intended users are not stated explicitly.
27c. Future research / generalizability	Future validation.	**Complete**	Independent cross-laboratory, protocol, sex/strain, severity, treatment, and biological-endpoint validation.

Checklist S2. Adapted PROBAST+AI risk-of-bias and applicability assessment

Study context	Entry
Study type	Three-classifier development and comparison with internal validation, followed by interpretation, a three-feature sensitivity analysis, equal-weight ensemble recovery-index construction, and surrogate analysis. PROBAST+AI use is adapted.
Target setting	Post-acquisition CatWalk XT gait analysis after T9 SCI.
Target population	Control Wistar rats from five source experiments.
Prediction target	Contemporaneous, experiment-specific Early/Middle/Late phase; IL4 day 1 is included in Early.
Analytical dataset	396 records from 115 animals: Early 175, Middle 116, Late 105.
Predictors	Prespecified 26-feature panel; timepoint, phase, animal, treatment, and experiment identifiers excluded.
Validation approach	5 × 5 animal-grouped nested cross-validation; 100-trial Optuna tuning per framework-fold.
Main models	Ordinal CatBoost, LightGBM, and XGBoost compared under the same resampling design; the final three-feature outputs were retained with equal weights by design.
Subsequent interpretation and ensemble analyses	Cross-fitted framework-specific calibration, TreeSHAP, ablation, three-feature classifier rerun, Platt-calibrated equal-weight ensemble recovery index, and GAM surrogate.
Intended use	Internal phase-associated gait analysis under the pooled protocol; not clinical or biological-recovery measurement.

Domain / assessment	Adapted judgment	Rationale
Participants: risk of bias	**Low to some concern**	Animal-grouped nested cross-validation prevents leakage; concerns are retrospective pooling, incomplete sex/age/weight metadata, and no independent validation.
Participants: applicability	**Some concern**	Limited to control Wistar rats and the represented CatWalk SCI protocol; other settings are untested.
Predictors: risk of bias	**Some concern**	Identifiers were excluded and transformations were fold-local; overall missingness was 40.6% and varied by phase, and bilateral aggregation removed side-specific information.
Predictors: applicability	**Low to some concern**	Definitions and code are available; use requires the canonical schema and comparable acquisition/processing.
Outcome: risk of bias	**Some concern**	Timepoint mapping is deterministic, but phase is a temporal proxy without an independent biological gold standard.
Outcome: applicability	**Some concern**	Supports contemporaneous phase classification, not prognosis, treatment response, or biological recovery.
Analysis: risk of bias	**Low to some concern**	Grouped nested cross-validation, fold-local engineering, out-of-fold estimation, clustered CIs, and leakage audits; concerns are 115 animals, internal validation, and post-selection analyses.
Analysis: applicability	**Some concern**	Targets new animals under the pooled protocol; experiment-blocked validation was not used because phase support differed structurally.
Overall: risk of bias	**Some concern**	Main concerns: proxy outcome, phase-dependent missingness, and no independent validation.
Overall: applicability	**Some concern**	Internally validated experimental framework; not an externally deployable or biological-recovery tool.

Checklist S3. ARRIVE 2.0 reporting checklist

ARRIVE 2.0 item	Status	Location / response
Study design	**Complete**	Secondary analysis of control-animal CatWalk records from five experiments; models compare recovery phase, not treatments.
Sample size	**Partly complete**	396 records/115 animals; no prospective ML sample-size calculation; all eligible records included.
Inclusion / exclusion criteria	**Complete**	Excluded baseline and rows with missing fp_stand_s; retained QC-passing controls with valid animal ID and phase (Table S2).
Randomisation	**Partly complete**	Source treatment randomisation was unavailable and not inferred. Analytical folds used seed 0, animal grouping, and experiment-phase balancing.
Blinding / masking	**Partly complete**	Recovery phase was deterministic and excluded from predictors. Paw prints were automatically classified, then corrected and validated by three researchers; masking was undocumented.
Outcome measures	**Complete**	Classification outcome: time-defined phase; predictors: 26 engineered features. Subsequent analytical outputs are the three-feature classifier panel, equal-weight ensemble recovery index, and GAM surrogate.
Statistical / analytical methods	**Complete**	Animal-grouped nested cross-validation; ordinal boosted trees; clustered inference; calibration; SHAP/ablation; recovery-index and GAM analyses.
Experimental animals	**Partly complete**	Wistar rats aged approximately 6–8 weeks and weighing 230–265 g. Final-cohort sex and individual age/weight were unavailable.
Experimental procedures	**Partly complete**	Main Methods: T9 clip injury, anesthesia/support, CatWalk acclimation/acquisition, paw-print review, and schedule; unavailable details were not inferred.
Housing and husbandry	**Partly complete**	Group housing reported; light/dark cycle, environment, and food/water records unavailable.
Ethical statement	**Complete (qualified)**	Source documentation identifies G-211/15 under German requirements; no broader claim.
Results: numbers analysed	**Complete**	396 records/115 animals: Early 175 records/93 animals; Middle 116/58; Late 105/33. Phase-specific animal counts overlap.
Results: adverse events / exclusions	**Partly complete**	Table S2 reports QC/cohort exclusions; source adverse events were unavailable.
Interpretation / scientific implications	**Complete**	Outputs are phase predictions and associations, not biological recovery, causality, or treatment effects.
Generalisability / translation	**Complete**	Estimand: new animals under the pooled protocol; external biological and cross-setting validation is needed.
Protocol registration	**Not registered**	No prospective registration; workflows and configurations are versioned.
Data and code availability	**Complete (restricted)**	Code: https://gitlab.com/sergienko.michael/sci. Aggregate outputs are available; raw and animal-level data are controlled.

Checklist S4. Data-leakage prevention checklist

Leakage-control item	Status	Implementation / evidence
Analytical unit defined before splitting	**Yes**	One row = one QC-retained CatWalk record.
Repeated-measures structure identified	**Yes**	Repeated records were clustered by animal.
Grouping variable isolated	**Yes**	factorized_animal was used for splitting, provenance, and inference—not as a predictor.
Outcome and time metadata excluded from X	**Yes**	Phase, timepoint, and cumulative-label variables were excluded.
Experiment and treatment metadata excluded from X	**Yes**	Experiment/group were used only for cohorting, balance, and summaries.
Frozen outer folds used	**Yes**	Five frozen outer folds were reused across frameworks, paired analyses, and the three-feature rerun.
Animal-level fold separation	**Validated**	Each animal occurred in one outer-test fold only.
Experiment-by-phase balance constrained	**Yes**	Experiment-phase balance was maintained where feasible subject to animal grouping.
Fixed seed and row lineage	**Yes**	Seed 0 and immutable row/animal provenance were stored.
Feature engineering inside pipelines	**Yes**	The feature transformer was fitted only on each active training partition.
Training-derived statistics isolated	**Yes**	Peak-loading means and SDs used training partitions only.
No full-dataset imputation or scaling	**Yes**	No global imputation or scaling; tree models handled missingness natively, and GAM preprocessing was fold-local.
Hyperparameter tuning nested	**Yes**	Optuna used grouped inner folds within each outer-training partition only.
Fold-specific final fitting	**Yes**	Pipelines were refitted on each outer-training partition; outer-test records were predicted once.
Performance based on OOF predictions	**Yes**	Performance used concatenated out-of-fold predictions.
Calibration cross-fitted	**Yes**	Calibrators excluded corresponding test records; the framework-specific rule retained no calibration for CatBoost.
Native TreeSHAP evaluated on held-out records	**Yes**	Native TreeSHAP used outer-test records; raw-margin additivity was checked.
Ablation separated from tuning	**Qualified**	Ablation refits used saved fold hyperparameters without re-optimization; this is a post-model diagnostic.
Post-selection panel explicitly qualified	**Qualified**	The three-feature panel was selected from out-of-fold evidence and underwent fresh nested cross-validation; it remains a post-selection sensitivity analysis.
Recovery index constructed from cross-fitted inputs	**Yes**	The ensemble index averages three Platt-calibrated three-feature scores with equal weights; each calibrator excluded the corresponding outer-test animal.
GAM target isolated	**Yes**	The GAM was trained to approximate the out-of-fold ensemble recovery index, not observed phase.
Full-data ensemble refits separated from evaluation	**Yes**	The three Platt-calibrated three-feature ensemble components were refitted after evaluation and excluded from performance estimates.
Validation claim bounded	**Yes**	Claims are limited to internal animal-separated validation under the pooled protocol.

Checklist S5. Reproducibility checklist

Reproducibility item	Status	Evidence / artifact
Software environment	**Yes**	Python 3.10.8; exact package versions are reported in Supplementary Table S18.
Dependency specification	**Yes**	Versioned project metadata and dependency lock.
Random seeds	**Yes**	Base seed 0; derived seeds and framework parameters stored.
Code availability	**Yes**	Repository: https://gitlab.com/sergienko.michael/sci
Source provenance	**Yes**	Workflow 00 manifest: five source experiments, row IDs, and hashes.
Configurations and run manifests	**Yes**	Each workflow records configuration, lineage, hashes, versions, seeds, counts, and status.
Cohort and lineage validation	**Yes**	Automated checks cover schema, provenance, row coverage, and reference counts.
Feature definitions	**Yes**	Supplementary Methods S3; Tables S4-S5: 19 inputs, aggregation, 26 formulas, epsilon, and fold statistics.
Fold assignments	**Yes**	Frozen animal-grouped fold assignments and summaries retained.
Optimization records	**Yes**	Optuna SQLite studies, trial histories, and best fold parameters retained.
Model and prediction artifacts	**Yes**	Fold models, out-of-fold probabilities, metrics, calibration/confusion outputs, and validation checks retained.
Interpretability artifacts	**Yes**	TreeSHAP/additivity, ablation predictions, consensus ranks, and correlations retained.
Three-feature panel and calibration artifacts	**Yes**	Three-feature definition, fresh nested cross-validation predictions, Platt outputs, and 26-versus-three-feature comparisons retained.
Recovery index and GAM artifacts	**Yes**	Framework/ensemble recovery indices, cross-fitted predictions, GAM, and fidelity metrics retained.
Classifier-ensemble artifact	**Yes**	The ensemble bundle contains all three Platt-calibrated three-feature components, equal one-third weights, schema metadata, a SHA-256 manifest, and a validated loader.
Data availability	**Restricted**	Aggregate outputs available; raw and animal-level data controlled.
External validation status	**Internal only**	No external dataset; claims are limited accordingly.