Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

A Unified Survival Benchmark for Temporal Dropout Risk Prediction in Learning Analytics

Rafael da Silva

rafael.dasilva@eastern.edu

Eastern University

Jeff Eicher

Eastern University

Gregory Longo

Eastern University

Research Article

Keywords: Learning Analytics, dropout prediction, temporal risk modelling, survival analysis, explainability, calibration

Posted Date: August 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10281741/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

A Unified Survival Benchmark for Temporal Dropout Risk Prediction in Learning Analytics

Rafael da Silva[1*], Jeff Eicher[1] and Gregory Longo[1]

[1*]Applied Data Science Program, Eastern University, St. Davids, PA, USA.

*Corresponding author(s). E-mail(s): rafael.dasilva@eastern.edu;
Contributing authors: jeff.eicher@eastern.edu; gregory.longo@eastern.edu;

Abstract

Although dropout remains a salient challenge in Learning Analytics, head-to-head model comparisons often rest on mismatched evaluation setups that emphasize ranking ability more than time-aware interpretability or calibration quality. Here we present a survival-oriented benchmark for temporal dropout risk modelling using the Open University Learning Analytics Dataset (OULAD). Two arms are compared: Family A: Dynamic Weekly, with models in person-period representation, and Family B: Static Early-Window, with an expanded roster of families: tree-based survival, parametric, and neural models. The evaluation protocol integrates four analytical layers: predictive performance, ablation, explainability, and calibration. Results are reported within each family separately, because a single numerical cross-family ranking would conflate genuine model differences with artifacts of temporal representation, to which survival metrics are known to be sensitive. Within Family B, Random Survival Forest showed the highest point estimates for time-dependent concordance and the lowest Brier scores across all three horizons; within Family A, Poisson Piecewise-Exponential showed the lowest point estimate for integrated Brier score within a tight five-model cluster. No-refit bootstrap resampling qualifies these positions as directional signals, not claims of strict superiority. Ablation and explainability analyses converged, across all models, on a shared finding: the dominant predictive signal was not primarily demographic or structural, but temporal and behavioral. Calibration corroborated this pattern in the better-discriminating models, except for XGBoost AFT, the sole outlier (analyzed in the Discussion). These results support unified, multi-dimensional benchmarking in Learning Analytics and situate dropout risk as a temporal-behavioral process rather than a function of static background attributes.

Keywords: Learning Analytics, dropout prediction, temporal risk modelling, survival analysis, explainability, calibration

1 Introduction

Student dropout poses a sustained challenge for higher education and for Learning Analytics as a field, driven not only by institutional and social costs, but by the difficulty of identifying risk patterns in time to support retention and persistence (Nagy & Molontay, 2023; Sandoval-Palis, Naranjo, Vidal, & Gilar-Corbí, 2020). Although this problem has attracted substantial research attention, a core tension persists: predictive models tend to be evaluated in terms of *who* is at risk, with less precision about *when* that risk intensifies across the academic trajectory (Adnan et al., 2021; Seo, Yang, Lee, & So, 2024).

The temporal dimension matters because dropout risk is not a static condition; it unfolds as engagement, activity, and academic progression accumulate week by week (Bañeres, Rodríguez-González, Guerrero-Roldán, & Cortadas, 2023; Mubarak, Cao, & Zhang, 2020). Yet the literature on dropout prediction still exhibits gaps at three interconnected levels. Many studies compare models under heterogeneous evaluation protocols, making it difficult to draw firm conclusions about differences across models (Coussement, Phan, De Caigny, Benoit, & Raes, 2020; Oqaidi, Aouhassi, & Mansouri, 2022). Even when comparative performance is strong in terms of discrimination, evaluations do not always incorporate dimensions essential for educational use: calibration and interpretability chief among them; calibration, in particular, refers to the numerical coherence between predicted risk probabilities and observed dropout rates (Nagy & Molontay, 2023). And the literature does not consistently translate performance differences into a substantive account of what kind of signal structures dropout risk: whether static student attributes or temporal and behavioral signals accumulated throughout the course (Seo et al., 2024; Zhidkikh, Heilala, Van Petegem, et al., 2024).

This paper responds to those gaps through a survival-oriented benchmark for temporal dropout risk modelling in Learning Analytics. The empirical basis is the Open University Learning Analytics Dataset (OULAD), a widely used resource in Learning Analytics and Educational Data Mining (Kuzilek, Hlosta, & Zdráhal, 2017). Its choice is deliberate rather than conventional: as the most-reused empirical testbed in the field, OULAD acts here as a standardized substrate whose familiarity allows performance differences to be read against a known baseline. The paper's contribution is not a new state-of-the-art score on this dataset; it is a demonstration that the same data yield different conclusions depending on how models are evaluated, which representational families they are organized into, and whether the evaluation protocol is unified across them. Data novelty is not the claim; evaluation architecture is. Under a unified evaluation framework—a single evaluation protocol (shared metrics, preprocessing, tuning, and infrastructure) applied across model families while preserving each family's methodological scope, not a merging or pooling of multiple datasets—two families are compared: Family A: Dynamic Weekly, with models in person-period representation, and Family B: Static Early-Window, with an expanded roster of families covering

tree-based survival, parametric, and neural approaches. The benchmark is designed to support within-family comparison under consistent evaluation protocols, making explicit the representational distinction between dynamic weekly and early-window approaches, under which a single numerical cross-family ranking would conflate model differences with artifacts of temporal representation. Crucially, the two families are two representations of the *same* underlying engagement signal, drawn from the same VLE logs; running them as parallel arms turns the representation choice into an experimental control, so any finding that holds across both families is robust to representation rather than an artifact of one encoding. The division is a control, not an arbitrary partition.

The benchmark is organized around four complementary dimensions. Predictive performance is evaluated through survival-oriented metrics under shared benchmark horizons: the Integrated Brier Score (IBS; a mean squared error integrated over the survival time axis) quantifies global probabilistic error, while time-dependent concordance (C-index) measures ordinal discriminative ability across the trajectory. An ablation analysis then assesses the extent to which performance depends on static feature blocks versus temporal and behavioral signals. An explainability layer identifies each family's dominant predictive drivers and compares the relative importance of feature blocks. A calibration layer is included because it is not enough for a model to rank cases well; predicted risks must also remain numerically coherent with observed outcomes, particularly when those estimates may inform support allocation and intervention decisions (Seo et al., 2024; Van Calster, McLernon, van Smeden, Wynants, & Steyerberg, 2019). Each layer earns its place: predictive performance asks *how well* models rank and predict; ablation asks *where* the signal comes from (static vs. temporal blocks); explainability asks *which* block dominates; and calibration asks whether the predicted probabilities are *numerically honest* enough to drive support decisions. The four are read jointly, not as interchangeable summaries.

Results are reported within each family separately, because a single numerical cross-family ranking would conflate genuine model differences with artifacts of temporal representation, to which survival metrics are known to be sensitive. Within Family B, Random Survival Forest showed the highest point estimates for time-dependent concordance and the lowest Brier scores across all three horizons; within Family A, Poisson Piecewise-Exponential showed the lowest point estimate for integrated Brier score within a tight five-model cluster. No-refit bootstrap sampling variability qualifies these positions as directional signals rather than absolute superiority (sampling variance only; see the Supplementary Material). More substantively, ablation and explainability converged on a consistent finding across all models: the dominant predictive signal was not primarily demographic or structural, but temporal and behavioral (Marcolino et al., 2025). Calibration corroborated this pattern in the better-discriminating models, except for XGBoost AFT, the sole outlier (analyzed in the Discussion). This convergence is what justifies maintaining two families rather than one: temporal-behavioral dominance appears in both representations and across linear, tree-based, and neural architectures—evidence no single family could provide.

The study contributes on four fronts: a unified evaluation framework that shares metrics, protocols, and infrastructure across two representational arms while preserving

the methodological separation that makes within-family comparisons valid; an argument that model comparison must extend beyond discrimination to calibration and interpretation; empirical evidence of temporal-behavioral signal dominance over static covariates; and a substantive reading of dropout risk as a structured temporal-behavioral process (Marcolino et al., 2025; Mubarak et al., 2020; Oqaidi et al., 2022).

Three research questions guide the article. **RQ1:** How do different survival-oriented models compare, under a unified evaluation protocol, in temporal dropout risk prediction? **RQ2:** To what extent does model performance depend on temporal and behavioral signals relative to static covariates? **RQ3:** How do explainability and calibration patterns help interpret, from a Learning Analytics perspective, what structures predicted dropout risk?

2 Literature Review

Dropout prediction occupies an established position within Learning Analytics, particularly in research on retention, persistence, and early identification of at-risk learners. Recent reviews of Predictive Learning Analytics in higher education confirm that retention and dropout are among the most frequently modelled outcomes, marking this as a well-developed line of inquiry (Sghir, Adadi, & Lahmer, 2022). This research agenda is closely tied to an institutional interest in using analytics not merely to describe learner behaviour, but to support timely decisions about intervention and student support (Sønderlund, Hughes, & Smith, 2018). Much of this empirical tradition has been built on widely reused datasets, with the Open University Learning Analytics Dataset (OULAD) serving as a recurring point of reference across both Learning Analytics and Educational Data Mining (Kuzilek et al., 2017).

A growing body of work recognizes that dropout risk should not be treated as a static binary outcome but as a process that emerges and evolves across the student trajectory. In online and blended settings, this risk can be observed through sequences of interaction, patterns of engagement, and shifts in academic behaviour, bringing dropout prediction closer to explicitly temporal formulations (Mubarak et al., 2020). The relevant question then becomes not only which students are at risk, but when that risk intensifies and at what point an intervention might be most useful. Early warning systems built around weekly windows or incremental course progression percentages reflect this shift, treating the timing of prediction as integral to its analytical value (Adnan et al., 2021; Akçapınar, Altun, & Aşkar, 2019). More recent work shows that both predictor importance and model performance can change over time from enrolment, reinforcing the view of dropout risk as a longitudinal phenomenon rather than a classification problem (Seo et al., 2024; Vaarma & Li, 2024). Correct enforcement of temporal observation boundaries (computing predictors exclusively from events prior to the prediction point, without drawing on post-window data) has been identified as a critical leakage-avoidance requirement for temporal LA models; its consequences for apparent predictive performance under standard evaluation protocols have been documented as a methodological concern that warrants explicit verification in benchmark designs (Kapoor & Narayanan, 2023).

Despite these advances, the comparative literature on dropout prediction still faces significant limitations. Studies vary considerably in target outcomes, data sources, feature types, algorithms, and evaluation metrics, producing a fragmented landscape in which direct comparison across findings is difficult (Sghir et al., 2022). Even explicitly comparative studies often operate under non-equivalent conditions (differing simultaneously in problem definition, temporal scope, feature availability, input representation, and experimental design). It therefore remains unclear whether reported performance differences reflect genuine advantages of one model over another or simply artefacts of the evaluation context. Coussement et al.'s systematic comparison across algorithms, for instance, is motivated precisely by this heterogeneity (Coussement et al., 2020). For Learning Analytics, the lack of unification undermines both the methodological cumulativeness of the field and the possibility of translating model findings into dependable guidance for educational practice (Sghir et al., 2022; Sønderlund et al., 2018).

A related limitation is the evaluation literature's strong centring on aggregate performance measures: accuracy, AUC, F1, or global error metrics (Coussement et al., 2020; Sghir et al., 2022). These are useful but insufficient. In Learning Analytics applications, correctly ranking students by risk is not enough; the risk estimates themselves must be coherent enough to support institutional decisions, follow-up prioritization, and resource allocation. Although the field does not always address calibration explicitly, several studies already treat individual risk probabilities as direct input for personalized intervention, making the numerical coherence of those predictions both analytically and practically consequential (Shiao et al., 2023; Xing & Du, 2019). More broadly, recent survival-methodology work has emphasized that different metrics capture different aspects of performance and should be interpreted jointly rather than as interchangeable summaries (Vasilev, Petrovskiy, & Mashechkin, 2023). The literature also stresses the importance of interpretability: when institutions need to understand the factors associated with risk and translate them into support strategies, discrimination alone is not enough. Work by Coussement et al. and Seo et al. illustrates the value of combining strong predictive performance with interpretation that reveals risk factors, segments, and behavioural patterns associated with dropout (Seo et al., 2024).

A deeper gap lies beneath these methodological concerns. Even when the literature identifies better-performing models, those gains are not consistently translated into an interpretation of what structures predicted dropout risk. Broad reviews of Predictive Learning Analytics continue to emphasize algorithms and performance metrics, with less attention to the substantive question of how model findings connect to risk mechanisms and educational response (Sghir et al., 2022; Sønderlund et al., 2018). One distinction is particularly consequential: the degree to which performance rests on relatively static student attributes versus temporal and behavioural signals derived from activity across the course. Recent studies confirm that this contrast carries empirical weight (Seo et al., 2024; Vaarma & Li, 2024), and it also changes the educational reading of risk, because structural variables and behavioural signals point to different kinds of institutional response. Landmarking approaches and dynamic prediction frameworks extend this idea to continuous-time settings by conditioning predictions on the patient's (or student's) history up to a landmark time, progressively updating

survival estimates as longitudinal observations accumulate; the present benchmark adopts a fixed early-window comparable representation as a methodological starting point, with time-varying covariate extensions reserved for future work.

These strands of literature converge on an unresolved need that is methodological as much as empirical: the field lacks sufficiently unified benchmarks that make model comparisons interpretable under a common protocol rather than across heterogeneous settings. The present study addresses that need by treating benchmark design itself as a contribution. Rather than proposing an exhaustive census of all survival architectures, it develops a survival-oriented benchmark structured around two families (Family A: Dynamic Weekly and Family B: Static Early-Window with an expanded roster of models), compared under shared evaluation conventions and under an explicit distinction between temporal representation regimes. It then uses ablation, explainability, and calibration to clarify not only which family performs best under the benchmark, but also what kind of information sustains that performance and what that means for Learning Analytics (Coussement et al., 2020; Sghir et al., 2022).

Table 1 positions representative studies from this literature against six methodological dimensions that structure the present benchmark. The dimensions are: **D1** — two representational families evaluated as an experimental control; **D2** — survival-specific metrics (IBS, time-dependent concordance) as the primary evaluation protocol; **D3** — OULAD used with weekly or daily VLE temporal granularity; **D4** — benchmark design treated as the primary contribution; **D5** — block-level feature ablation assessed with survival metrics; **D6** — equity or fairness analysis by student subgroup. The table reveals that no prior study satisfies D1 (dual representational families as experimental control) and only one partially addresses D5 (block ablation with survival metrics), which are precisely the dimensions the present benchmark is designed to fill.

Table 1 Positioning of representative prior studies against six methodological dimensions (D1–D6) of the present benchmark. **D1**: dual representation families as experimental control; **D2**: survival metrics as primary protocol; **D3**: OULAD with weekly/daily VLE granularity; **D4**: benchmark design as primary contribution; **D5**: block-level ablation with survival metrics; **D6**: equity/fairness analysis by subgroup. Symbols: ✓ = present and central; ∼ = partial; – = absent.

Study	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**
Gutierrez-Pachas et al. (2023) (Gutierrez-Pachas et al., 2023)	–	✓	–	∼	–	∼
Pan et al. (2022) (Pan et al., 2022)	–	✓	–	–	–	–
Adnan et al. (2021) (Adnan et al., 2021)	–	–	∼	–	–	–
He et al. (2020) (He et al., 2020)	–	–	✓	–	–	–
Villar & de Andrade (2024) (Villar & de Andrade, 2024)	–	–	–	∼	–	–
Coussement et al. (2020) (Coussement et al., 2020)	–	–	–	∼	–	–
Fazil & Albahlal (2025) (Fazil & Albahlal, 2025)	–	–	✓	–	∼	–
Gao & Ni (2025) (Gao & Ni, 2025)	–	–	∼	–	–	✓
Vaarma et al. (2024) (Vaarma & Li, 2024)	–	–	–	–	–	✓
Zanellati et al. (2024) (Zanellati et al., 2024)	–	–	–	–	–	✓
This work	✓	✓	✓	✓	✓	∼

Study framework.

Two commitments organize the paper: (i) dropout is treated as a *temporal-behavioral process*, not a static background trait; and (ii) consequently, *benchmark design itself* is the contribution—a unified, multi-dimensional protocol organized by representational family—rather than a single best-model claim. Every methodological choice below follows from these two.

3 Methods

To address the unification and interpretability gaps identified above (heterogeneous evaluation protocols, neglected calibration, and insufficient signal attribution), the benchmark applies consistent evaluation protocols across both families while respecting the methodological boundaries between survival model types. Two representational regimes structure the benchmark throughout this paper. **Family A: Dynamic Weekly** uses a *person-period* data representation (one row per enrollment per week), in which weekly conditional hazards are estimated and then accumulated into enrollment-level survival curves; this family is also referred to as the weekly discrete-time family. **Family B: Static Early-Window** uses a fixed *enrollment-level* representation built from early-window summaries, in which a single feature vector per enrollment feeds directly into continuous-time survival models. These two family labels and their corresponding data representations are used interchangeably throughout the paper and always refer to the same methodological distinction. The two-family design provides a shared evaluation infrastructure (consistent metrics, preprocessing, tuning, and reporting conventions) while respecting the representational distinction between the two families. Figure 1 provides a structured overview of the full benchmark pipeline. Two orientation tables anchor the rest of the paper: Table 4 situates all 14 models by family, model type, and inductive bias, and Table 5 gives a plain-language reading of every metric and when to trust it. Readers can use them as a map before the per-layer detail.

This within-family reporting choice is grounded in the sensitivity of survival metrics to temporal representation. Varying only the discretization or representation scheme shifts the integrated Brier score (Kim, 2023; Kvamme & Borgan, 2021) and produces non-monotonic variation in time-dependent concordance for one and the same model (Nooraldaim, Xue, Lai, Zhu, & Wang, 2026); more broadly, different survival metrics capture different aspects of performance and are not interchangeable (Park, Park, Kim, & Park, 2021; Vasilev et al., 2023). Because Family A and Family B operate under distinct temporal representations, a single numerical cross-family ranking would conflate genuine model differences with artifacts of temporal representation. Results are therefore reported within each family, over shared enrollment-level horizons, rather than as a cross-family ranking. The cross-family ranking is deliberately *withheld* (not infeasible): it would conflate model differences with representation artifacts, whereas convergent within-family findings are precisely what the two-arm control is designed to expose.

Table 2 Feature-block membership across the two families. The overlap is *intentional* and confined to the shared *static structural* block; the *exclusive* block is what creates each family's representational regime. Because both families derive from the same OULAD source, agreement across them is not confounded by differing data. Full variable lists are in Table 3.

Shared (both families)	Family A only (dynamic weekly)	Family B only (static early-window)
Static structural block: demographics and prior academic record (OULAD `studentInfo`).	*Dynamic temporal-behavioral* block: week-by-week engagement signals on the person-period panel.	*Early-window behavioral* block: first-weeks engagement compressed into enrollment-level scalars.

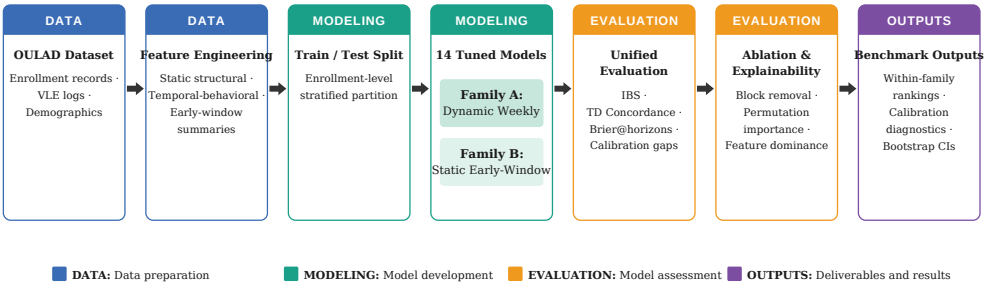

Figure 1 Overview of the full benchmark pipeline, organized into four color-coded phases. The **Data** phase (blue) covers raw data ingestion and feature engineering from OULAD sources into static, temporal-behavioral, and early-window feature blocks. The **Modeling** phase (teal) encompasses the enrollment-level train/test partition and the training of 14 tuned models across the two representational arms (Family A: Dynamic Weekly; Family B: Static Early-Window). The **Evaluation** phase (orange) applies the unified assessment protocol — IBS, time-dependent concordance, Brier scores at three horizons, and calibration gaps — followed by ablation and explainability analyses. The **Outputs** phase (purple) consolidates within-family rankings, calibration diagnostics, and bootstrap uncertainty estimates as the final benchmark deliverables.

3.1 Dataset

The empirical basis for the benchmark is the *Open University Learning Analytics Dataset* (OULAD), which integrates demographic, academic, assessment, and virtual learning environment (VLE) interaction data (Kuzilek et al., 2017). The dataset is publicly available at https://research.stem.open.ac.uk/ouanalyse/dataset/. Direct download at http://schools.stem.open.ac.uk/cdn/files/anonymisedData.zip. Its continued use across Learning Analytics and Educational Data Mining studies supports its suitability as a benchmark dataset for comparative methodological work (Mihăescu & Popescu, 2021).

The dataset links administrative and behavioural information: student demographics, course registration records, assessment data, and large-scale traces of online VLE

interaction (Kuzilek et al., 2017). The central analytical unit is the *enrollment*, defined as the student–module–presentation combination. This unit corresponds to the level at which withdrawal is administratively observed and allows the benchmark to represent both weekly risk trajectories and enrollment-level summaries of early engagement (Prenkaj, Velardi, Stilo, Distante, & Faralli, 2020; Rodríguez, Villanueva, Dombrovskaia, & Valenzuela, 2023).

The benchmark dataset comprises 32,593 enrollments spanning 7 modules and 4 presentations. The observed event rate (enrollments with *Withdrawn* status and a valid unregistration date) is 22.7%, with the remainder right-censored at the end of their observation window.

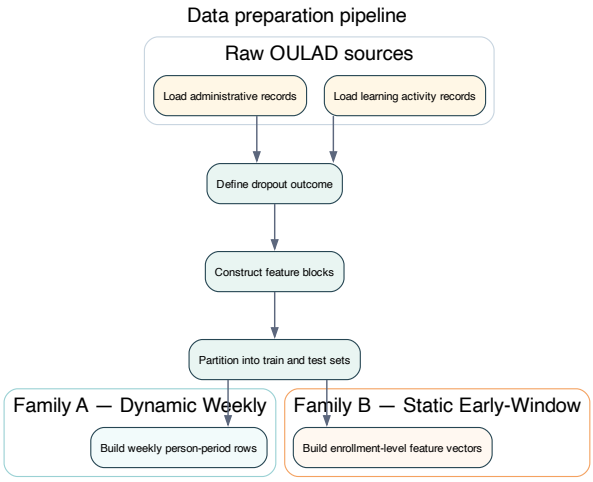

Figure 2 Data preparation pipeline from raw OULAD sources to benchmark-ready datasets for the discrete-time and static early-window families.

3.2 Outcome definition and prediction task

The outcome of interest is student dropout, operationalized through the administratively recorded status *Withdrawn* for enrollments with a valid unregistration date. This follows standard practice in dropout modelling, where administrative status transitions, non-reenrolment, or institutional withdrawal records define the target event (Cannistrà, Masci, Ieva, Agasisti, & Paganoni, 2021; Rodríguez et al., 2023).

For each enrollment i, let T_i denote the observed time in weeks and let $\delta_i \in \{0, 1\}$ denote the event indicator, where $\delta_i = 1$ indicates an observed withdrawal and $\delta_i = 0$ indicates censoring. Formally,

$$\delta_i = \begin{cases} 1, & \text{if enrollment } i \text{ has observed withdrawal,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Rather than treating dropout as a static end-of-course binary label, the study formulates the task as *temporal dropout risk prediction*, consistent with work that treats student dropout as a longitudinal or time-dependent process (Mduma, Kalegele, & Machuve, 2019; Prenkaj et al., 2020). Time is discretized at the weekly level, a granularity that is both practically meaningful and widely adopted in educational early-warning settings (Adnan et al., 2021; Aljohani, Fayoumi, & Hassan, 2019). Under this design, the benchmark compares models in terms of how they represent temporal dropout risk under common evaluation horizons, not merely whether they predict eventual withdrawal.

3.3 Data preparation and feature engineering

Data preparation followed two parallel branches, each designed to accommodate a different temporal representation under a common benchmark framework.

For the discrete-time models, the benchmark adopts a *person-period* representation in which each enrollment is expanded into sequential weekly observations until the event or censoring. This structure follows discrete-time event history modelling, where the conditional hazard is estimated across repeated intervals for each observational unit (Allison, 1982; Suresh, Severn, & Ghosh, 2022). The weekly format supports dynamic temporal features (current activity, recency, streak, and cumulative behavioural summaries), consistent with educational research showing the value of clickstream intensity and evolving activity traces in modelling academic risk (Adnan et al., 2021; Waheed et al., 2020).

For the continuous-time comparable models, the benchmark uses an enrollment-level representation with *early-window summaries*. Temporal information from the early part of the course is compressed into a fixed vector of aggregated features capturing initial engagement and behavioural participation. This design aligns with survival modelling frameworks that operate on fixed covariates while predicting time-to-event risk (Suresh et al., 2022; Wiegrebe, Kopper, Sonabend, & Bender, 2023), and it allows the continuous-time models to remain methodologically comparable to the discrete-time families while preserving a conceptually distinct view of temporal information. This representation does not constitute a general continuous-time formulation with time-dependent covariates; Family B operates on fixed enrollment-level representations derived from the early observation window. As a consequence, differences in performance between the two families partly reflect this representational constraint and should not be interpreted as evidence about the general superiority of continuous-time over discrete-time survival modelling. Landmarking and joint modelling approaches constitute a natural extension of this design to settings where time-varying covariates are available; these represent an explicit direction for future comparative benchmarking.

Features were organized into three conceptually meaningful blocks: (1) static structural covariates, including demographic and prior academic information; (2) dynamic temporal-behavioural signals, derived from week-by-week engagement; and (3) early-window behavioural summaries, used by the static early-window branch.

This block-based structure supports the ablation, explainability, and substantive interpretation analyses that follow (Wiegrebe et al., 2023). Table 3 lists all variables in each block together with their computation method.

Preprocessing was standardized across models as far as their input representations allowed. Numeric variables were imputed with the median and categorical variables with an explicit missing category. Categorical predictors were encoded through one-hot expansion with unknown categories ignored at test time, and numeric variables were standardized. Crucially, all preprocessing transformations were fitted on the training partition only and then applied to validation and test data, so that no information from held-out observations influenced feature transformation.

Class imbalance handling was intentionally conservative: no explicit resampling or reweighting was applied across any of the 14 tuned families. A sensitivity experiment confirmed that explicit inverse-frequency class weighting enlarged rather than reduced Family A calibration gaps (weighted IBS 0.44–0.52, calibration gap 0.24–0.30 at horizon 10, versus unweighted IBS 0.14, gap 0.09–0.10), validating the unweighted design choice.

Family A uses only person-period information observed up to horizon t; Family B generates a full survival curve from fixed early-window inputs (main window = first four weeks), preventing use of post-window information. A sensitivity grid ($w \in \{2, 4, 6, 8, 10\}$ weeks) was evaluated across window lengths; results are reported at the canonical window ($w = 4$ weeks), which was held fixed for all primary benchmark comparisons (Supplementary Table S10).

See Figure 2 for a schematic summary of the dual data preparation pipeline.

3.4 Models

The benchmark compares 14 tuned models organized across Family A and Family B. This methodological breadth is deliberate: conclusions about hierarchies or dominant predictive signals are more defensible when they emerge consistently across families with structurally different inductive biases (linear, tree-based, parametric, and neural) than when they rest on a single representative. Family A: Dynamic Weekly comprises five models operating on the person-period representation: (1) a linear discrete-time hazard model, (2) a neural discrete-time survival model, (3) a Poisson piecewise-exponential model, (4) a gradient-boosted weekly hazard model, and (5) a CatBoost weekly hazard model. Family B: Static Early-Window comprises nine models operating on fixed enrollment-level early-window representations: (1) a comparable Cox model, (2) DeepSurv, (3) Random Survival Forest, (4) Gradient-Boosted Cox, (5) Weibull AFT, (6) Royston-Parmar, (7) XGBoost AFT, (8) Neural-MTLR, and (9) DeepHit. Table 4 provides a structured overview of all 14 families organized by representational group, model type, and inductive bias. The equations that follow describe the mathematical anchors of each family; all other families share the same input representation as their respective family anchor.

The first anchor family is a *linear discrete-time hazard model*, implemented on the person-period dataset. For enrollment i at week t, the discrete-time hazard is defined as

$$h_i(t) = P(T_i = t \mid T_i \geq t, \mathbf{x}_{it}), \tag{2}$$

where $\mathbf{x}_{it}$ denotes the covariates observed for enrollment i at week t. In the linear specification, the weekly conditional hazard is linked to the predictors through

$$\operatorname{logit}\big(h_i(t)\big) = \alpha_t + \mathbf{x}_{it}^\top \beta, \tag{3}$$

where α_t is a time-specific intercept and β is the coefficient vector (Suresh et al., 2022).

The enrollment-level survival curve induced by the weekly hazards is then reconstructed as

$$S_i(t) = P(T_i > t \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}) = \prod_{\tau=1}^{t} \big(1 - h_i(\tau)\big), \tag{4}$$

with fixed-horizon event risk obtained by complementarity,

$$r_i(t) = P(T_i \leq t \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}) = 1 - S_i(t). \tag{5}$$

The Family A's neural anchor is a *neural discrete-time survival model*, also implemented on the person-period representation. Instead of assuming a linear relationship between covariates and the interval hazard, this family uses a neural architecture to model the weekly hazard flexibly across intervals. Conceptually, this replaces the linear predictor in Equation 3 with a nonlinear function $f_\theta(\mathbf{x}_{it})$, allowing more complex interactions while preserving interval-based time-to-event prediction (Kvamme & Borgan, 2021; Wiegrebe et al., 2023).

$$h_i(t) = \sigma(\alpha_t + f_\theta(\mathbf{x}_{it}))\,, \tag{6}$$

where $\sigma(z) = \big(1 + e^{-z}\big)^{-1}$ is the logistic link.

The Family B's linear anchor is a *comparable Cox model*, estimated at the enrollment level using the early-window summaries. The Cox proportional hazards model specifies the hazard as

$$\lambda_i(t \mid \mathbf{x}_i) = \lambda_0(t) \exp(\mathbf{x}_i^\top \beta), \tag{7}$$

where $\lambda_0(t)$ is the baseline hazard and $\mathbf{x}_i$ is the fixed covariate vector for enrollment i (Cox, 1972).

The Family B's neural anchor is *DeepSurv*, also estimated at the enrollment level. DeepSurv extends the Cox model by replacing the linear predictor with a neural network, yielding

$$\lambda_i(t \mid \mathbf{x}_i) = \lambda_0(t) \exp\big(f_\theta(\mathbf{x}_i)\big), \tag{8}$$

where $f_\theta(\mathbf{x}_i)$ is the nonlinear representation learned by the network (Katzman et al., 2018; Kvamme, Borgan, & Scheel, 2019).

The primary comparative focus is on tuned versions, in which models are compared under systematically optimized rather than default parameterizations (Kvamme et al., 2019).

3.5 Unified evaluation protocol

All tuned models were assessed under a unified survival-oriented evaluation protocol designed to reduce interpretive ambiguity across differing problem definitions, temporal scopes, and metric selections (Burk et al., 2024).

Evaluation was organized around a hierarchy of complementary survival-oriented metrics. The main metrics include the *Integrated Brier Score* (IBS), *time-dependent concordance*, horizon-specific Brier scores, and horizon-specific calibration at the shared benchmark horizons of weeks 10, 20, and 30. For a horizon t, let $\hat{G}(t)$ denote the Kaplan–Meier estimate of the censoring survival function. The IPCW Brier score used in the benchmark may be expressed as

$$\mathrm{BS}^{\mathrm{IPCW}}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{I}(T_i \le t,\, \delta_i = 1)}{\hat{G}(T_i^-)}\hat{S}_i(t)^2 \;+\; \frac{\mathbb{I}(T_i > t)}{\hat{G}(t)}\big(1-\hat{S}_i(t)\big)^2\right], \tag{9}$$

where $\hat{S}_i(t)$ is the predicted survival probability for enrollment i at horizon t. The integrated version reported throughout the article is

$$\mathrm{IBS} = \frac{1}{\tau_{\max}}\int_0^{\tau_{\max}} \mathrm{BS}^{\mathrm{IPCW}}(u)\,du, \tag{10}$$

where $\tau_{\max}$ is the upper evaluation horizon. In the empirical implementation, horizon-specific Brier scores and IBS were computed under right-censoring through inverse-probability-of-censoring weighting (IPCW) using a Kaplan–Meier estimate of the censoring distribution, following the evaluation convention implemented through `pycox`. IBS summarizes prediction error over the evaluation interval and served as the main global measure of probabilistic performance (Gerds & Schumacher, 2006; Graf, Schmoor, Sauerbrei, & Schumacher, 1999; Park et al., 2021).

Discrimination was evaluated through *time-dependent concordance* in the formulation of Antolini et al. (Antolini, Boracchi, & Biganzoli, 2005), which assesses concordance of predicted survival functions with observed event ordering across the full event-time horizon, rather than through an unspecified static-risk C-index. This distinction matters because the benchmark compares full survival trajectories across common horizons, not merely fixed scalar risk scores. Accordingly, this article refers to this quantity as a time-dependent concordance measure, even when compact tables retain the conventional shorthand *C-index* (Harrell, Califf, Pryor, Lee, & Rosati, 1982; Longato, Vettoretti, & Di Camillo, 2020; Park et al., 2021).

The protocol distinguishes dynamic from static early-window prediction settings: discrete-time families update weekly hazards sequentially using only week-t information, accumulating enrollment-level survival as in Equations 4–5; static early-window families generate a full survival curve from fixed early-window inputs. Both branches report at shared enrollment-level horizons without information leakage.

The enrollment-level split was stratified by event status with no identity leakage; a contextual audit confirmed that curricular context (all 7 modules, 4 presentations, 22 module-presentation combinations) was fully shared across train and test. Findings

generalize across enrollments under shared curricular context, not to unseen module or presentation settings.

Horizon-specific calibration was evaluated through a bin-based risk calibration procedure. At each benchmark horizon h, predicted event risk was grouped into quantile-based bins $b = 1, \ldots, B$, and calibration error was summarized as the sample-size-weighted mean absolute gap

$$\mathrm{Calib}(h) = \sum_{b=1}^{B} \frac{n_b}{n} \left| \bar{r}_b(h) - \tilde{y}_b(h) \right|, \tag{11}$$

where n_b is the size of bin b, $\bar{r}_b(h)$ is the mean predicted event risk in that bin, and $\tilde{y}_b(h)$ is the IPCW-adjusted empirical event rate at horizon h. The empirical event rates used in this comparison were therefore estimated under right-censoring adjustment, consistent with the IPCW framework applied to the Brier score and IBS computations, rather than as unadjusted observed proportions. These reported calibration values are therefore horizon-specific weighted absolute calibration gaps. In addition, the strengthened calibration audit estimated approximate calibration intercept and slope by horizon through a weighted fit on the reliability-bin summaries, so that calibration could be interpreted jointly through primary gap, supporting Brier, support, and logit-scale slope/intercept diagnostics rather than through a single scalar alone (Park et al., 2021).

Because leading margins were numerically small, uncertainty was assessed through enrollment-level bootstrap resampling (200 resamples, no model refit) on the held-out test set. Frozen predictions were resampled to obtain empirical intervals for IBS, time-dependent concordance, and Brier horizons; these reflect sampling variability only and are lower bounds on total uncertainty; retraining-based intervals, which would propagate model estimation variance, could be substantially wider. For time-dependent concordance, RSF's leading position was preserved in a large majority of resamples; for IBS, the bootstrap provided a directional signal only.

The comparable Cox model was subjected to a formal proportional-hazards audit yielding a broad-departure-from-proportionality classification (10 of 38 covariates, 26.3%); results from the comparable Cox branch should be narrated as from an approximate specification. DeepSurv shares the Cox-type structure but is not amenable to classical Schoenfeld diagnostics; detailed scope boundaries are reported in the Supplementary Material.

Table 5 summarises all evaluation metrics used in the benchmark, organised by analytical role, with a plain-language description to support readers less familiar with survival-specific evaluation conventions.

This multi-metric design reflects the recognition that no single metric can adequately summarize model quality in temporal risk prediction (Park et al., 2021).

See Figure 3 for a schematic overview of the unified benchmark evaluation flow. Additional protocol details, calibration-by-horizon diagnostics, bootstrap uncertainty summaries, and proportional-hazards audit outputs are reported in the Supplementary Material.

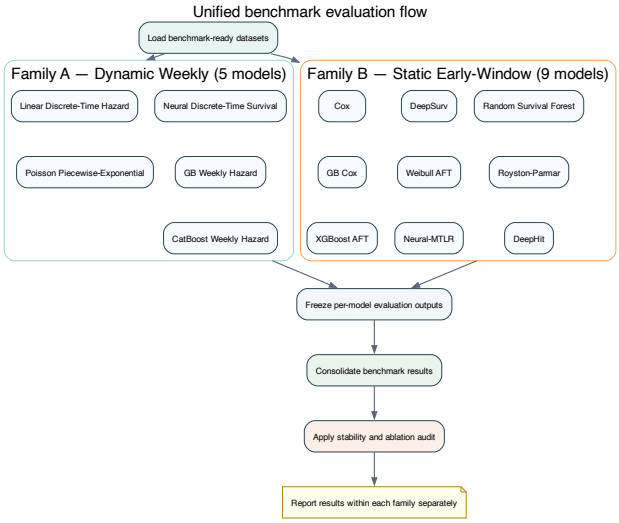

Figure 3 Unified benchmark evaluation flow used to compare the models under a common survival-oriented protocol.

3.6 Ablation design

The ablation analysis estimates the relative contribution of major feature blocks to model performance. Rather than interpreting performance at the level of the full model alone, the benchmark evaluates tuned variants in which selected predictor groups are removed or isolated, making it possible to assess how far predictive performance depends on static structural covariates versus temporal and behavioural information.

Ablation serves here as both a sensitivity analysis and an interpretive tool. Comparing full-feature models with variants that exclude specific blocks reveals which kinds of signal most strongly sustain predictive performance (Li & Janson, 2024).

3.7 Explainability and calibration overview

The benchmark includes two complementary interpretive layers: explainability and calibration.

Explainability methods were matched to each model. For intrinsically interpretable models, global interpretation draws on model coefficients and hazard-related effect sizes (Cox, 1972). For the nonlinear families, grouped permutation importance provides a model-agnostic measure of the global contribution of individual predictors and conceptually defined feature blocks (Fisher, Rudin, & Dominici, 2019). Four blocks structure this analysis: *static structural covariates* (demographic and prior academic background, applicable across all families); *dynamic temporal-behavioral signals* (week-by-week engagement and activity traces, Family A only); *discrete time index* (the weekly time step as a standalone predictor of conditional hazard, Family A only; in

the linear discrete-time model, the time step is encoded as week-specific intercepts α_t in Equation 3, whereas in the neural and gradient-boosted families it enters as an ordinal covariate in $\mathbf{x}_{it}$; the block encompasses both roles as predictors of the weekly conditional hazard); and *early-window behavioral summaries* (compressed early-course engagement aggregates, Family B only). This block-based organization enables comparison of explanatory patterns across paradigms while anchoring interpretation in the same feature taxonomy used by the ablation layer.

Calibration provides an explicit check on probabilistic coherence. Where discrimination measures whether students are correctly ranked by relative risk, calibration evaluates whether numerical risk estimates remain coherent with observed event frequencies over the benchmark horizons. In the present benchmark, this layer is interpreted through a primary horizon-wise weighted absolute calibration gap together with supporting slope/intercept diagnostics derived from the reliability-bin summaries (Park et al., 2021).

Because the present study evaluates calibration as a robustness layer rather than as a model-updating study, post-hoc recalibration is treated as a future strengthening step rather than as a procedure applied to the held-out benchmark comparison itself. Reliability diagrams (calibration plots) are included in the Supplementary Material (Supplementary Figures S1–S4) as visual complements to the tabulated calibration gaps; recalibration experiments (isotonic regression, beta calibration, or Platt scaling applied to the frozen held-out predictions) are deferred to future work. The Family A calibration gaps (0.089–0.098 at horizon 10, increasing to 0.123–0.149 at horizon 30) are the primary candidate for recalibration follow-up, given their systematically larger magnitude relative to the best Family B models.

See Figure 4 for a schematic overview of how explainability and calibration are integrated into the benchmark.

3.8 Reproducibility and implementation

The analytical pipeline was implemented as a reproducible notebook-based workflow, with systematic export of all tables, figures, and metadata reported in the study. This design ensures traceability across data preparation, preprocessing, model fitting, evaluation, ablation, explainability, calibration, and paper integration (Peng, 2011). The complete pipeline (including all stage scripts, frozen model artifacts, tables, and figures) is publicly available at https://github.com/rafa-rodriguess/dropout_benchmark. Every number in the paper is generated by a single export step from these versioned, frozen artifacts, so each reported value traces deterministically to a pipeline output rather than to an intermediate notebook state, a raw database extraction, or an upstream script.

Each of the 14 tuned models followed the same discipline: preprocessing design, enrollment-level validation partitioning, bounded search scope, selection criterion, and complexity controls. Individual records are consolidated in the supplementary audit (Supplementary Table S2).

Tuning was deliberately controlled rather than exhaustive: the goal was stronger, comparable benchmark representatives, not unbounded per-family optimization. All models were subjected to an equivalent tuning budget of 24 candidate evaluations, ensuring that no model benefited from a disproportionately larger search effort. The budget

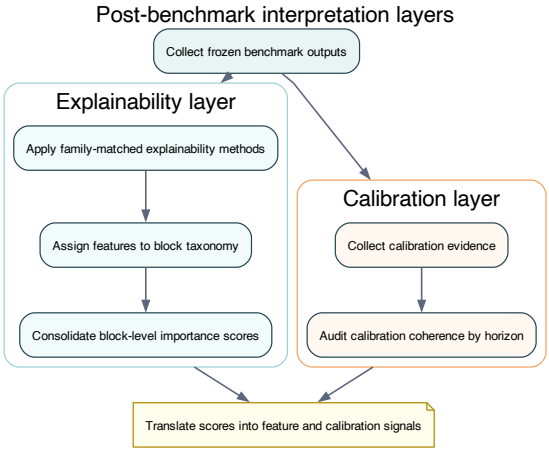

Figure 4 Overview of the explainability and calibration layers used as complementary post-benchmark analyses across the models.

of 24 candidates was set by the most complex model in the benchmark (DeepSurv), whose architecture and regularisation space required the largest grid; all remaining models were tuned under the same ceiling to ensure symmetric search effort.

Use of AI-assisted tools

During the preparation of this work, the authors used large language model tools (Claude, ChatGPT) to assist with structural revision and editorial drafting of selected sections, with the aim of improving clarity and readability. All outputs were reviewed and edited by the authors, and the authors take full responsibility for the content of this publication. Use was limited to writing assistance; all benchmark design decisions, data analysis, model implementation, and interpretation were performed exclusively by the authors.

4 Results

4.1 Main benchmark comparison

Reading guide.

If you read one line per family: within **Family B**, Random Survival Forest leads on time-dependent concordance (0.674) and on all three Brier horizons, with XGBoost AFT as the sole outlier; within **Family A**, five models sit inside a 0.0011 IBS band, with Poisson Piecewise-Exponential narrowly ahead. The detail below supports these two lines; the orientation tables are Table 4 (models) and Table 5 (metrics).

Results are reported within each family separately. The two families use different risk formulations (weekly person-period hazards versus fixed early-window survival predictions), so that a single combined numerical ranking would conflate genuine model differences with artifacts of temporal representation, to which metrics such as the IBS are known to be sensitive. Within each family, results are organized around the within-family leader—Poisson Piecewise-Exponential for Family A and Random Survival Forest for Family B—with the remaining models reported as a comparative roster (Table 6); the systematic ablation is carried out on a representative eight-model subset, four per family. The full 14-model roster is load-bearing for the main comparison precisely because the within-family spread is itself the result (a 0.0011 IBS band in Family A; 0.1118–0.1190 for the eight non-outlier models in Family B): a sparse roster could not establish the full landscape of model performance, while the eight-model subset is the representative cross-section used for signal attribution.

Within Family B, *Random Survival Forest* showed the highest point estimate for time-dependent concordance (0.674) and the lowest Brier scores across all three horizons; the remaining seven models ranged from 0.1131 to 0.1190 on IBS, with *XGBoost AFT* as the sole outlier (IBS 0.150, TD concordance 0.577). Within Family A, all five models clustered within a 0.0011 IBS band (0.1399–0.1410), with *Poisson Piecewise-Exponential* showing the lowest point estimate for integrated Brier score.

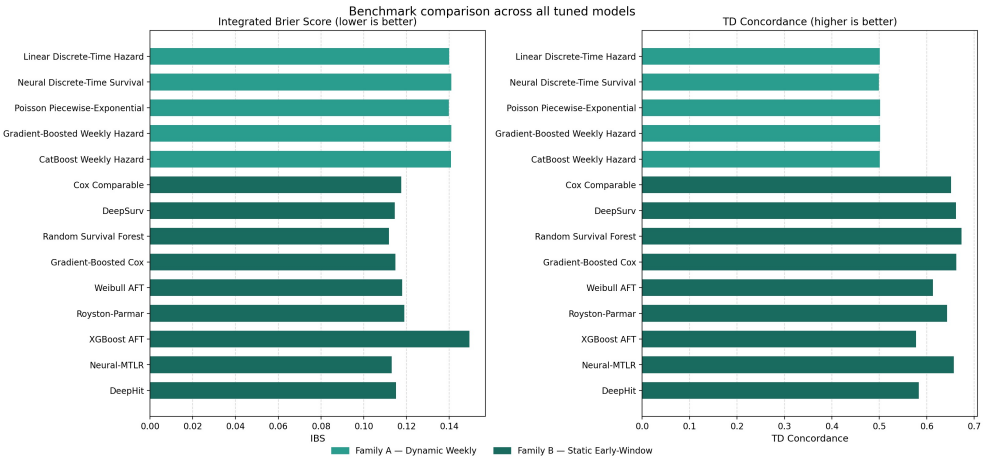

Figure 5 Benchmark comparison of all 14 models under the unified survival-oriented protocol, grouped by family. Within Family B, RSF leads on time-dependent concordance and all Brier horizons; XGBoost AFT is a visible outlier. Within Family A, all five models cluster tightly on IBS and horizon-specific Brier; Brier values for Family A are reconstructed from accumulated weekly hazards.

Within Family B, the RSF–DeepSurv gap on TD concordance was 0.012; the Brier@10 gap between RSF and the next-best model was 0.001. RSF showed the best point estimate on all reported metrics within that family; these differences may not be statistically significant, as the no-refit bootstrap reflects sampling variability only. The no-refit bootstrap showed RSF leading time-dependent concordance in the majority of

resamples and maintaining Brier-horizon leadership under resampling; IBS differences among the top eight Family B models were modest, with RSF leading at 0.1118 and the next-best cluster ranging from 0.1131 to 0.1190. Within Family A, *Linear Discrete-Time Hazard* showed the lowest Brier@10 (0.1195) and *Poisson Piecewise-Exponential* showed the lowest Brier@20 and Brier@30 (0.1621 and 0.1931); all five models fell within a 0.0011 IBS band and a 0.0018 Brier@10 band, with no single model dominating across all metrics simultaneously.

The proportional-hazards audit on the comparable Cox model assigned a global classification of broad departure from proportionality: 10 of 38 tested covariates (26.3%) showed evidence of possible non-proportionality at $\alpha = 0.05$. No other Family B model was subjected to an identical classical proportional-hazards diagnostic.

The train–test partition was leakage-free at the enrollment identity level; a contextual audit confirmed complete overlap of modules, presentations, and module-presentation combinations across splits.

4.2 Ablation results: static versus temporal-behavioral signal

A systematic ablation experiment was conducted on eight of the fourteen tuned models: four from Family A (*Poisson Piecewise-Exponential*, *Linear Discrete-Time Hazard*, *GB Weekly Hazard*, and *Neural Discrete-Time Survival*) and four from Family B (*Cox Comparable*, *DeepSurv*, *RSF*, and *Neural-MTLR*), removing each feature block in turn while holding all other modelling choices fixed. The eight models span linear, gradient-boosted tree, and neural architectures across both families, ensuring that the finding on temporal signal dominance does not rest on a single modelling paradigm or a single model. For Family B, IBS-based and concordance-based ablation are both reported: ablation is performed at the enrollment level under the same fixed early-window representation used for the primary benchmark, so each feature-removal variant yields a well-defined enrollment-level survival curve that can be scored with IBS under the IPCW estimator. For Family A, only concordance-based ablation is reported: model predictions are person-period hazard scores evaluated row-by-row across weekly intervals, whereas IBS is defined over enrollment-level survival curves reconstructed from those hazards; computing IBS under each feature-removal variant would require re-running the full survival-curve reconstruction pipeline for every ablated model, which was outside the scope of the current benchmark freeze; concordance-based ablation captures the ranking impact of feature removal under the family's native discrimination objective and is methodologically consistent with the family's evaluation contract.

Within Family B, all four models showed a larger IBS increase when the early-window behavioral block (compressed early-course engagement aggregates, including clickstream activity counts and session summaries; see Explainability layer in Methods) was removed than when static covariates (demographic and prior academic background variables) were removed; on average, temporal signal removal increased IBS by 0.0083 versus 0.0051 for static removal, with IBS ratios ranging from 1.20 (RSF) to 2.46 (Cox Comparable). Across all eight models, concordance decreased more when temporal signal was withheld than when static covariates were removed, with Family A temporal removal reducing concordance by 0.081–0.102 and static removal causing smaller changes across all four Family A models.

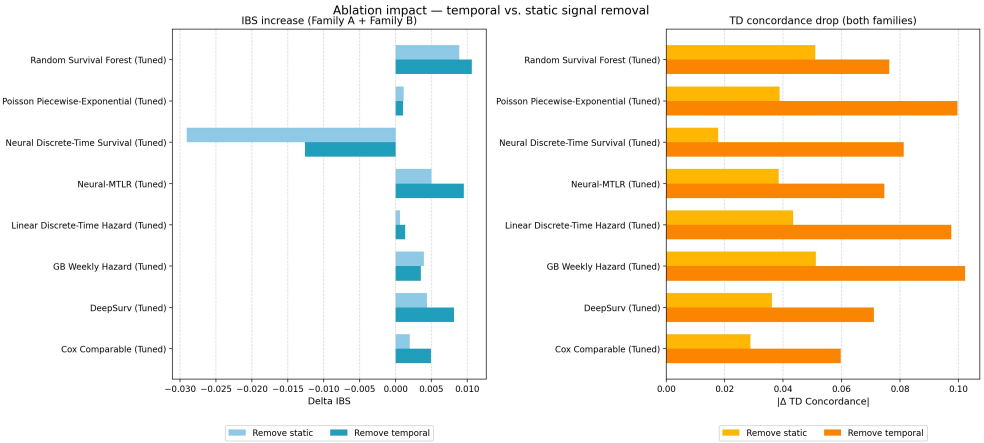

Figure 6 Ablation impact across eight models from both families, contrasting the loss associated with removing static versus temporal-behavioral signal. Across all families, removing temporal-behavioral signal is more damaging than removing static covariates. The IBS panel (left) shows only the four Family B models because IBS-based ablation requires enrollment-level survival curves, which are directly available for the early-window Family B but not for Family A, where predictions are weekly hazard scores. The TD concordance panel (right) includes all eight models.

4.3 Explainability across models

The paper-facing explainability export covers eight tuned models spanning both families: four from Family A (*Poisson Piecewise-Exponential*, *Linear Discrete-Time Hazard*, *GB Weekly Hazard*, and *Neural Discrete-Time Survival*) and four from Family B (*Cox Comparable*, *DeepSurv*, *RSF*, and *Neural-MTLR*). These models provide complementary inference across linear, gradient-boosted tree, and neural architectures in both the weekly person-period and early-window enrollment-level representations. In all eight models, the dominant feature *block* is temporal: Family A models show either *dynamic_temporal_behavioral* or *discrete_time_index* as the dominant block; Family B models uniformly show *early_window_behavior* as dominant. The top individual driver is a temporal or early-window variable in seven of the eight models; in *GB Weekly Hazard*, the highest individual-feature attribution corresponds to a static covariate (*studied_credits*), yet the dominant block at the aggregate level remains *discrete_time_index*, consistent with the temporal dominance pattern across the full ensemble. This distinction between individual-feature attribution and block-level dominance reflects the dual role of the *discrete_time_index* block, which encompasses both time-specific intercepts and ordinal time-step covariates depending on the model.

Ablation and explainability converge on the same pattern: temporal or early-window signal was dominant across all eight representative models, regardless of family or architecture.

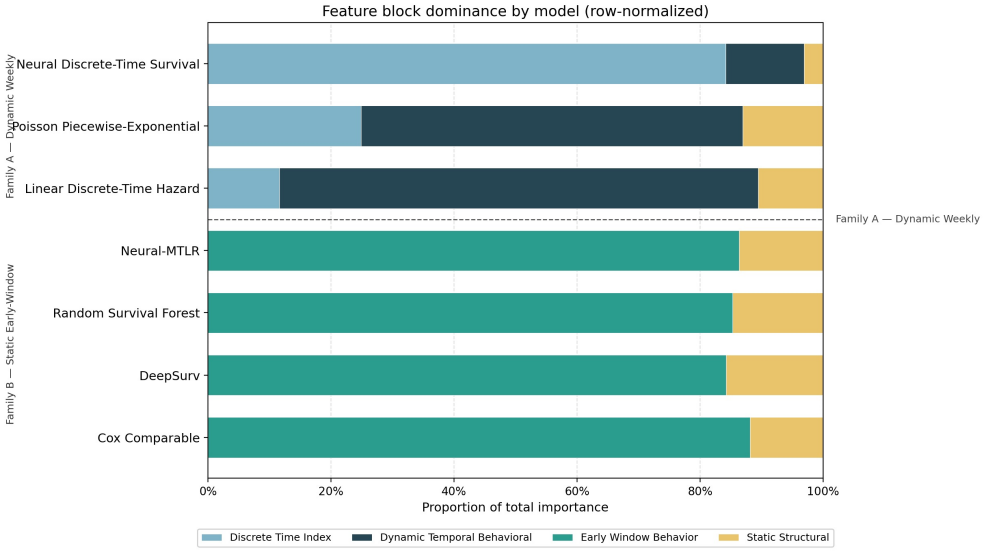

Figure 7 Normalized explainability block dominance across eight models from both families included in the paper-facing export. Values are normalized within model to highlight relative block importance; the figure should be read as a within-model comparison of dominant signal blocks, not as an absolute cross-model magnitude comparison.

4.4 Calibration results

Calibration was evaluated through a quantile-bin procedure on predicted event risk at the benchmark horizons, using IPCW-adjusted empirical event rates in each bin. The reported values are the horizon-wise weighted absolute gaps defined in Equation 11. This metric is the primary calibration criterion; slope and intercept diagnostics from a weighted reliability-bin fit are treated as strengthening auxiliaries, not as primary outputs.

Calibration results for both families are reported in Table 10. For Family B, predictions are generated from fixed enrollment-level representations and horizon risk is extracted directly from the predicted survival curve. For Family A, weekly hazard predictions are accumulated into a survival curve as in Equation 4, and fixed-horizon cumulative risk is extracted as in Equation 5; the evaluation uses study-level enrollment durations so that students who do not withdraw remain evaluable as survivors, preserving the full test set across all enrollment windows.

Within Family B, *Random Survival Forest* showed the best calibration at horizon 10 and the most consistently strong calibration profile across all three horizons (gaps: 0.012, 0.016, 0.014); most other models showed moderate calibration with gaps ranging from 0.011 to 0.044 (excluding Neural-MTLR, which was wider). *XGBoost AFT* was a clear exception, with calibration gaps of 0.124, 0.178, and 0.218 at horizons 10, 20, and 30, substantially larger than any other Family B model. *DeepSurv* and *Cox Comparable* showed the lowest gaps at horizons 20 and 30, respectively; *Neural-MTLR* showed wider gaps across all horizons. Within Family A, all five models showed calibration gaps

between 0.089 and 0.098 at horizon 10, with *Gradient-Boosted Weekly Hazard* lowest at 0.089; gaps increased with horizon, ranging from 0.123 to 0.149 at horizon 30. Family A calibration gaps are larger in absolute magnitude than those of the best Family B models, but they remain meaningfully below the XGBoost AFT outlier at all horizons.

A class-weighting sensitivity experiment (Methods) confirmed the unweighted protocol as the better-calibrated design choice. A horizon-level comparison of calibration gaps and reliability diagrams are presented in the Supplementary Material.

5 Discussion

5.1 Benchmark hierarchy and methodological scope

Within Family B, RSF showed consistently the best point estimate across all reported metrics. This profile was maintained under bootstrap resampling: RSF led time-dependent concordance in 90% of resamples and maintained Brier-horizon leadership; RSF also held a 64% IBS rank-1 share, a directional signal that should not be read as strict dominance given overlapping confidence intervals—Neural-MTLR captured 27% of rank-1 IBS draws under resampling, indicating the IBS hierarchy is partially shared between the two top models. The within-family hierarchy is therefore largely stable rather than perfectly rigid. For practitioners deploying early-warning systems, consistent multi-metric leadership across concordance and Brier horizons provides a more reliable basis for allocating support resources than single-metric optimisation; RSF's profile across all four metrics, maintained under bootstrap resampling, makes it the more actionable choice within Family B. Within Family A, the tight IBS band means no single model holds a stable bootstrap lead.

Two scope conditions bound the interpretation of these results. The proportional-hazards audit classified the comparable Cox model as broad departure from proportionality (10 of 38 covariates showing possible non-proportionality at $\alpha = 0.05$), establishing that this anchor is an approximate specification rather than a fully assumption-clean Cox model; all results from this model should be read with this approximation in scope. The formal PH audit applies only to the comparable Cox anchor and should not be read as a symmetric diagnostic pass for all Family B models. Additionally, the benchmark evaluates generalization across enrollments under shared curricular context—modules, presentations, and module-presentation combinations overlap completely across splits—not transportability to unseen curricular settings.

The result that *Random Survival Forest* led on time-dependent concordance and horizon-specific Brier scores within Family B suggests that non-parametric flexibility was sufficient to capture the dominant enrollment-level patterns in the early-window representation, even relative to the neural static early-window models tested here. The two families are best read as complementary methodological evidence about the same dropout risk phenomenon, not as a ranked comparison of weekly versus early-window modelling strategies.

5.2 Feature signal attribution

The ablation experiment localizes the source of predictive performance. The main gains were not sustained by relatively stable background attributes; the most severe losses emerged when models were deprived of variables representing activity, progression, and engagement across the course. In this dataset, dropout risk does not appear primarily as a background trait but as a pattern that becomes visible through the temporal trajectory of academic and behavioral participation (Mubarak et al., 2020; Vaarma & Li, 2024). That engagement features constructed as primary indicators do, in fact, dominate is a necessary validity check rather than a surprising structural claim: it confirms the pipeline captures the intended temporal signal and is not driven by noise or by an unintended proxy. Read this way, the result is a control that licenses the subsequent cross-model comparison, not a deflation of it. Such a reading aligns with a Learning Analytics perspective in which prediction's analytical value lies not only in classification but in surfacing temporally meaningful signals of persistence, disengagement, and possible withdrawal (Alalawi, Athauda, & Chiong, 2024; Ifenthaler & Yau, 2020).

The top individual drivers for the Family B models are behavioral summaries from the early observation window: *clicks_first_4_weeks* for Cox Comparable and *active_weeks_first_4* for DeepSurv, RSF, and Neural-MTLR. These variables summarize observable early engagement that covaries with dropout risk, but do not by themselves support causal claims about why dropout occurs (Arnold et al., 2020; Hooshyar & Yang, 2024).

These ablation and explainability findings are paradigm-agnostic: all eight models assigned dominance to a temporal block regardless of family or architecture, with both analyses converging on the same directional conclusion across linear, tree-based, and neural paradigms (Fisher et al., 2019).

5.3 Calibration coherence and risk structure

Risk models can produce strong discrimination while generating unreliable probability estimates (Kvamme et al., 2019; Van Calster et al., 2019). In the present benchmark, calibration reinforced rather than challenged the main benchmark hierarchy: within Family B, *RSF* showed the best calibration at horizon 10 and the most consistently strong calibration profile across all three horizons (gaps: 0.0121, 0.0162, 0.0138), confirming that its discrimination advantage is not accompanied by calibration coherence loss. The *XGBoost AFT* outlier also aligns with its benchmark outlier status on IBS and TD concordance, suggesting that its overall weaker result reflects a consistent estimation failure rather than an isolated metric artefact. From a risk-structure perspective, the miscalibration of XGBoost AFT's time-invariant AFT parameterisation while more flexible hazard formulations (RSF, DeepSurv, Cox) remain better-calibrated suggests that dropout hazard ratios in this dataset are temporally heterogeneous; the miscalibration is therefore a structural diagnostic about the risk process, indicating that dropout risk does not unfold with time-constant covariate acceleration, and not merely a basis for model selection. Within Family A, all five models produced calibration gaps below 0.10 at horizon 10, substantially smaller than the XGBoost AFT

outlier, confirming that the Family A weekly hazard predictions, when read as horizon-specific cumulative risk estimates, are numerically coherent with observed dropout rates. Family A gaps are larger in absolute magnitude than those of the best Family B models, though substantially below the XGBoost AFT outlier.

5.4 Synthesis

Benchmark, ablation, explainability, calibration, and bootstrap evidence converged on a consistent interpretation. The main comparative conclusions do not rest on a single metric or on a single reading of performance. Rather, they emerge from the alignment among global probabilistic error, time-dependent discrimination, horizon-specific error, and calibration coherence (evaluated within each family separately, and supported by enrollment-level bootstrap resampling as a ranking-stability diagnostic).

The answers to the three research questions reinforced rather than contradicted each other, with one important qualification for each. For RQ1, the benchmark yielded a largely stable within-family hierarchy: within Family B, *Random Survival Forest* led on time-dependent concordance (0.674; 90% rank-1 bootstrap share, largely stable) and on all three Brier horizons; within Family A, *Poisson Piecewise-Exponential* led narrowly on IBS (0.1399) within a tight five-model cluster; IBS bootstrap leadership was directional (RSF rank-1 share 64%; Neural-MTLR 27%), though intervals still overlap and the ranking should not be treated as strictly dominant. The ordering within each family is consistent but not strictly dominant: both families show modest absolute margins. Family A and Family B cannot be pooled into a single global ranking without erasing the representational difference between weekly discrete-time hazard and fixed early-window continuous-time survival predictions.

For RQ2, the systematic ablation on eight representative models (four per family) showed that temporal-behavioral signal removal caused substantially larger performance loss than static removal in all Family B models (IBS ratio range: 1.20–2.46) and on concordance across all eight models; this finding holds regardless of modelling paradigm.

For RQ3, the paper-facing explainability export (covering eight models across both families) found that the dominant feature block in all eight models is temporal (*dynamic_temporal_behavioral*, *discrete_time_index*, or *early_window_behavior*), a result that now converges with rather than complements the ablation evidence. Calibration within Family B showed that *RSF* had the best calibration profile (gaps: 0.012–0.016), most other models showed moderate gaps (0.011–0.044, excluding Neural-MTLR, which was wider), with *XGBoost AFT* as a consistent extreme outlier across all metrics; Family A models showed calibration gaps between 0.089 and 0.098 at horizon 10, confirming they generate numerically coherent probability estimates rather than degenerate risk scores. Together, explainability and calibration converge on a consistent answer to RQ3: dropout risk in this dataset is structured as a temporally evolving process driven primarily by behavioral engagement signals, with a risk function amenable to well-calibrated survival modeling when the chosen formulation is sufficiently flexible to accommodate time-varying hazard dynamics.

Methodologically, these results suggest that comparisons between temporal dropout risk models in Learning Analytics may benefit from unified, multi-dimensional protocols, as opposed to isolated metrics or cross-study synthesis under heterogeneous

settings (Farhood, Joudah, Beheshti, & Müller, 2024). This also places the paper in direct dialogue with recent survival-evaluation arguments that caution against single-metric reading and instead recommend evaluation layers that remain explicit about what each metric validates and what it does not (Lillelund, Qi, Greiner, & Pedersen, 2025). Substantively, dropout risk in this context appears as a process structured by temporal signal (early participation, activity intensity, weekly progression, and recent disengagement), and this conclusion is supported consistently by both ablation and explainability analyses, with both converging on temporal dominance across all eight representative models and both families (Marcolino et al., 2025).

The leading position of Random Survival Forest in Family B is consistent with findings from neutral survival benchmarks on non-educational datasets. Kvamme et al. (Kvamme et al., 2019) evaluated RSF, Cox, DeepSurv, and DeepHit across several standard clinical survival datasets and found RSF to be persistently competitive, though model leadership varied by dataset and scoring rule. Burk et al. (Burk et al., 2024) conducted a large-scale neutral comparison of survival methods on low-dimensional data and likewise found tree-based ensembles to be competitive with the Cox baseline, particularly in datasets with moderate to strong signal-to-noise ratios. The consistency between the present OULAD result and these external references strengthens the interpretation that RSF's discrimination advantage reflects algorithmic characteristics rather than dataset-specific artifact. That said, model rankings in those neutral benchmarks varied across datasets and metrics, indicating that replication of the present benchmark on other educational datasets remains a necessary step before generalizing the Family B hierarchy beyond the shared-context OULAD setting.

All findings are bounded by shared curricular context (all modules and presentations appear in both splits), and bootstrap intervals reflect enrollment-sampling variance only, not a formal hypothesis test or retraining-based generalization claim.

Takeaway.

In one line: under a unified protocol, 14 models across two representations agree that dropout is a temporal-behavioral process, and within the early-window family the tree-based Random Survival Forest leads on a multi-metric basis (TD concordance 0.674, IBS 0.1118) with a 90% bootstrap rank-1 share on discrimination. What the benchmark does *not* establish: transportability to unseen modules/presentations, or formal hypothesis tests of the rankings (the no-refit bootstrap bounds sampling variance only). For a practitioner: prefer RSF within the early-window family on the strength of consistent multi-metric leadership, not a single score.

5.5 Practical implications

Three conclusions from the benchmark carry a direct operational reading for institutions that collect LMS engagement data.

First, early-window prediction is viable at week four. RSF trained on the first four weeks of engagement consistently outperformed alternatives on discrimination and calibration. An institution with access to click counts and weekly activity logs in its VLE can produce calibrated dropout-risk scores at week 4, before the critical withdrawal window in most modules. Because RSF's probabilities are well-calibrated (Family B

gaps: 0.012–0.016), a threshold such as $P(\text{dropout}) > 0.3$ corresponds to a meaningful risk level, not an arbitrary score rank. A practical three-step protocol follows from these properties: score all active enrolments at the end of week 4 using the first-four-week behavioral summary; flag those with $\hat{P}(\text{dropout}) > 0.30$ for proactive tutor contact; and, among flagged students, prioritise those with simultaneously low VLE activity and few active weeks—the dominant SHAP contributors across all Family B models. The lower IBS of RSF translates directly into this threshold being closer to the empirical withdrawal rate, making 0.30 operationally meaningful rather than a percentile cut.

Second, the dominant predictive signal is early engagement, not background attributes. Ablation and explainability results agree: what predicts dropout is observable in-course behavior—clicks in the first four weeks and number of active weeks—rather than demographic background or prior qualifications. Institutions do not need complex data pipelines; the most predictive variables are the simplest to collect from any LMS.

Third, weekly monitoring adds value under a different deployment contract. Family A models can, in principle, update risk scores throughout the course rather than only at week 4. The trade-off is weaker enrollment-level calibration relative to Family B. A practical hybrid would use an early-window RSF score at week 4 to trigger initial outreach, then use weekly engagement trends as a qualitative signal for ongoing case management, without treating weekly model scores as calibrated probabilities.

One constraint applies across all three readings: the benchmark was conducted under shared curricular context, and deployment on unseen modules or different institutional settings requires local revalidation before the reported metrics can be treated as expected performance bounds (Sønderlund et al., 2018).

6 Conclusion

This study developed and applied a unified survival-oriented benchmark for temporal dropout risk modelling in Learning Analytics, comparing 14 tuned models organized into two methodologically distinct arms: a Family A of five weekly discrete-time hazard models and a Family B of nine early-window continuous-time survival models. Its purpose was not only to compare models, but to make those comparisons methodologically clearer and educationally interpretable under a shared evaluation protocol.

Within Family B, *Random Survival Forest* led on time-dependent concordance (0.674) and on all three reported Brier horizons, with bootstrap evidence supporting a largely stable rank-1 position on discrimination across 90% of enrollment-level resamples. *RSF* showed the best calibration profile within Family B; *XGBoost AFT* was a consistent extreme outlier across all metrics. Within Family A, the five models clustered within a 0.0011 IBS band, with *Poisson Piecewise-Exponential* narrowly leading. The two families use different risk formulations and evaluation footprints, so that a single cross-family ranking would conflate genuine model differences with representational artifacts rather than measure real differences. A systematic ablation on eight representative models (four per family) showed that temporal-behavioral signal removal caused larger performance loss than static removal in all Family B models (IBS ratio range: 1.20–2.46) and on concordance across all eight models. The paper-facing explainability export (covering eight models across both families) identified a temporal

or early-window block as dominant in all eight models, a finding that converges with rather than merely complements the ablation evidence.

Two contributions stand out. Methodologically, the study shows the value of comparing temporal dropout risk models under a unified, multi-dimensional protocol organized by family group, rather than through isolated metrics or indirect synthesis across heterogeneous studies (Collins et al., 2024; Dwivedi, Joshi, Laurent, Bengio, & Bresson, 2022). Substantively, the ablation and explainability results converge on a paradigm-agnostic conclusion: temporal-behavioral signal (whether captured as weekly activity in Family A or early-window engagement summaries in Family B) dominates both the aggregate-removal impact and the block-level attribution across all eight representative models and both families. The benchmark structure itself is the transferable contribution: the OULAD provided a standardised substrate for its demonstration, not an end in itself.

Limitations

Three scope qualifications bound the interpretation of all reported findings: (i) external generalizability is constrained by the single-dataset design; (ii) bootstrap confidence intervals are no-refit lower bounds on ranking uncertainty, reflecting enrollment-sampling variance only and not the additional estimation variance from model refitting or hyperparameter re-selection; and (iii) post-hoc recalibration was not applied to the held-out predictions.

The benchmark is based on a single dataset and should not be treated as automatically generalizable to other institutional settings. The split evaluated generalization under shared curricular context (all 7 modules and 4 presentations appear in both partitions; Supplementary Table S5) rather than module- or presentation-level transportability, and bootstrap intervals reflect enrollment-sampling variance only. The early-window choice ($w = 4$) was validated against a sensitivity grid ($w \in \{2, 4, 6, 8, 10\}$ weeks; Supplementary Table S10). Eight directions remain as future work: decision-analytic evaluation (decision curves, net benefit); fairness and subgroup-performance auditing; comparison to non-survival LA/EDM baselines (e.g., logistic regression, decision trees); cross-institutional replication; recalibration experiments for Family A models (reliability diagrams in Supplementary Figures S1–S4); cross-module transportability studies (Alalawi et al., 2024; Sønderlund et al., 2018); extension to time-varying covariate designs via landmarking; and ensemble combinations across models.

Table 3 Feature blocks used in the benchmark, with computation method. Static structural covariates are drawn directly from the OULAD `studentInfo` table. Dynamic temporal-behavioral features are aggregated from `studentVle` at the enrollment×week level; derived features (*italic*) are computed on the person-period panel. Early-window behavioral summaries collapse the first four weeks of the dynamic panel into enrollment-level scalars. Block membership governs the ablation and grouped permutation importance analyses.

Block / Variable	**Description**	**Computation**
Static structural covariates — all models		
`gender`	Student gender	Direct from `studentInfo`
`region`	UK region of residence	Direct from `studentInfo`
`highest_education`	Highest prior qualification	Direct from `studentInfo`
`imd_band`	Index of Multiple Deprivation decile	Direct from `studentInfo`
`age_band`	Age group at registration	Direct from `studentInfo`
`disability`	Declared disability status	Direct from `studentInfo`
`num_of_prev_attempts`	Prior module attempts	Direct from `studentInfo`
`studied_credits`	Credits enrolled in module	Direct from `studentInfo`
Dynamic temporal-behavioral — Family A only		
`total_clicks_week`	VLE clicks in current week	`SUM(sum_click)` from `studentVle`, by enrollment×week
`n_vle_rows_week`	VLE interaction rows this week	`COUNT(*)` from `studentVle`, by enrollment×week
`active_this_week`	Any VLE activity this week (binary)	1 if any `studentVle` row exists for that enrollment×week
`n_distinct_sites_week`	Distinct VLE sites visited this week	`COUNT(DISTINCT id_site)` from `studentVle`, by enrollment×week
`cum_clicks_until_t`	Cumulative VLE clicks up to week t	Running `SUM(total_clicks_week)` over expanding window ordered by week
`recency`	Weeks since last active week	0 if active this week; $t - t_{\text{last active}}$ otherwise; $t + 1$ if never active
`streak`	Consecutive active weeks up to t	Row number within current uninterrupted active run; 0 if inactive
Early-window behavioral summaries — Family B only (weeks 0–3)		
`clicks_first_4_weeks`	Total VLE clicks in weeks 0–3	`SUM(total_clicks_week)` for week < 4, grouped by enrollment
`active_weeks_first_4`	Active weeks count in weeks 0–3	`SUM(active_this_week)` for week < 4, grouped by enrollment
`mean_clicks_first_4_weeks`	Mean clicks per active week in weeks 0–3	`clicks_first_4_weeks / active_weeks_first_4`; 0 if never active

Table 4 Overview of the 14 models organized by representational group (Family A/B), model type, and inductive bias. Models within each group are ordered from linear to neural, with tree-based and parametric families in between.

Group	Model type	Model	Inductive bias
Family A	Linear	Linear Discrete-Time Hazard	Linear
		Poisson Piecewise-Exponential	Linear (log-rate)
	Neural	Neural Discrete-Time Survival	Neural
	Gradient-Boosted	GB Weekly Hazard	Tree ensemble
		CatBoost Weekly Hazard	Tree ensemble
Family B	Cox	Cox Comparable	Linear (PH)
	Neural Cox	DeepSurv	Neural (PH)
	Parametric	Weibull AFT	Parametric (AFT)
		Royston-Parmar	Parametric (spline PH)
	Tree ensemble	Random Survival Forest	Tree ensemble
		Gradient-Boosted Cox	Tree ensemble (PH)
		XGBoost AFT	Tree ensemble (AFT)
	Neural	Neural-MTLR	Neural
		DeepHit	Neural

Note. **Family A: Dynamic Weekly** ≡ person-period representation (one row per enrollment per week; weekly hazards accumulated into enrollment-level survival curves). **Family B: Static Early-Window** ≡ fixed enrollment-level representation (one row per enrollment; early-window summaries fed directly into continuous-time survival models). The two family labels and their corresponding data representations are used interchangeably throughout the paper.

Table 5 Reference summary of all evaluation metrics used in the benchmark. IBS is the survival analogue of the Mean Squared Error (MSE) used in classical ML regression: where MSE averages squared prediction errors on a fixed outcome, IBS averages squared errors on the predicted survival probability over the full event-time axis, adjusted for censoring via IPCW. For error-based metrics, lower scores correspond to stronger performance, whereas discrimination metrics improve as scores increase. Calibration gaps nearer zero reflect closer agreement between predicted risk and observed dropout rates.

Metric	Role	Better when	What it measures
Primary performance metrics			
IPCW weight $\hat{G}(t)^{-1}$	Censoring correction	—	Inverse-probability-of-censoring weight, estimated non-parametrically via Kaplan-Meier on the censoring process. Upweights uncensored observations to correct for the selective removal of censored students from the risk set as follow-up progresses. Applied internally to Brier@h, IBS, and Calib@h; not a standalone evaluation metric.
Brier@10, @20, @30	Error	Lower ↓	Horizon-specific probabilistic error: squared difference between predicted survival probability and observed event status at a fixed week, averaged over enrollments with IPCW adjustment. Analogous to MSE at a snapshot in time.
Integrated Brier Score (IBS)	Error	Lower ↓	Global probabilistic error integrated over the full evaluation window. Survival analogue of MSE: where MSE averages squared errors on a fixed outcome, IBS averages squared errors on the predicted survival curve over time, adjusted for censoring.
TD Concordance (C-index)	Discrimination	Higher ↑	Time-dependent concordance (Antolini): probability that a student who drops out earlier receives a higher predicted risk than one who drops out later or does not drop out. Measures ranking quality across the full event-time trajectory, not just at a fixed horizon.
Calibration metrics			
Calib@10, @20, @30	Calibration	Lower ↓	Horizon-specific weighted absolute gap between mean predicted event risk and IPCW-adjusted observed dropout rate within quantile bins. Zero means predicted probabilities match observed rates exactly at that horizon.
Calibration slope & intercept	Calibration (auxiliary)	Slope ≈ 1, intercept ≈ 0	Logit-scale linear fit on reliability-bin summaries. Slope < 1 indicates overconfidence (predictions too extreme); slope > 1 indicates underconfidence; intercept $\neq 0$ indicates systematic over- or underestimation. Treated as strengthening diagnostics, not primary ranking criteria.
Ranking stability diagnostic			
Bootstrap rank-1 share / 95% CI	Stability	Higher share = more stable	Proportion of 200 enrollment-level resamples in which a model holds rank 1 on a given metric (frozen predictions, no model refit). Captures sampling variability in ranking, not estimation variance. Intervals are lower bounds on total uncertainty.

Table 6 Main benchmark comparison across all 14 models, grouped by family. For Family A, Brier-horizon scores are computed from enrollment-level survival curves reconstructed from accumulated weekly hazards, using the same IPCW methodology as Family B; both families are evaluated on the full metric set. Lower IBS and Brier scores indicate better probabilistic accuracy (lower mean squared survival error); higher time-dependent concordance indicates better discrimination.

Model	IBS	TD Concordance	Brier@10	Brier@20	Brier@30
Family A: Dynamic Weekly					
Poisson Piecewise-Exponential	0.1399	0.5019	0.1213	0.1621	0.1931
Linear Discrete-Time Hazard	0.1400	0.5013	0.1195	0.1637	0.1971
CatBoost Weekly Hazard	0.1408	0.5011	0.1203	0.1644	0.1980
Neural Discrete-Time Survival	0.1410	0.4992	0.1199	0.1648	0.1990
Gradient-Boosted Weekly Hazard	0.1410	0.5018	0.1206	0.1646	0.1981
Family B: Static Early-Window					
Random Survival Forest	0.1118	0.6735	0.0947	0.1329	0.1574
Neural-MTLR	0.1131	0.6569	0.0954	0.1340	0.1588
DeepSurv	0.1145	0.6619	0.0978	0.1343	0.1589
Gradient-Boosted Cox	0.1148	0.6621	0.1000	0.1361	0.1601
DeepHit	0.1151	0.5835	0.0973	0.1363	0.1614
Cox Comparable	0.1176	0.6512	0.1012	0.1378	0.1623
Weibull AFT	0.1180	0.6131	0.1016	0.1380	0.1624
Royston-Parmar	0.1190	0.6430	0.1035	0.1398	0.1641
XGBoost AFT	0.1495	0.5772	0.1241	0.1773	0.2174

Table 7 Ablation results: Family B (early-window continuous-time survival). IBS ratios above 1.0 indicate greater dependence on temporal-behavioral signal than on static covariates.

Model	ΔIBS static	ΔIBS temporal	ΔTD static	ΔTD temporal	IBS ratio
Cox Comparable	0.0020	0.0049	−0.0288	−0.0598	2.46
DeepSurv	0.0044	0.0082	−0.0361	−0.0711	1.86
Neural-MTLR	0.0050	0.0095	−0.0385	−0.0747	1.90
Random Survival Forest	0.0089	0.0106	−0.0510	−0.0764	1.20

Table 8 Ablation results: Family A (weekly discrete-time hazard). Only concordance-based evidence is reported (ΔTD concordance). IBS-based ablation was not computed for Family A models: the per-person-period hazard evaluation surface differs in temporal footprint from the enrollment-level survival curve on which IBS is defined, and converting the ablation to IBS would require re-running the full curve-reconstruction pipeline under each feature-removal variant. Concordance-based ablation is therefore the methodologically consistent metric for this family.

Model	ΔTD static	ΔTD temporal
Gradient-Boosted Weekly Hazard	−0.0513	−0.1023
Linear Discrete-Time Hazard	−0.0434	−0.0976
Neural Discrete-Time Survival	−0.0178	−0.0812
Poisson Piecewise-Exponential	−0.0389	−0.0997

Table 9 Cross-family explainability summary for eight models across both families. *Early-window behavioral*: compressed early-course engagement summaries. *Dynamic temporal-behavioral*: week-by-week activity features. *Discrete time index*: the time-step covariate (week number).

Model	Top driver	Dominant block
Family B		
Cox Comparable	clicks_first_4_weeks	early_window_behavior
DeepSurv	active_weeks_first_4	early_window_behavior
Random Survival Forest	active_weeks_first_4	early_window_behavior
Neural-MTLR	active_weeks_first_4	early_window_behavior
Family A		
Linear Discrete-Time Hazard	n_vle_rows_week	dynamic_temporal_behavioral
Poisson Piecewise-Exponential	n_vle_rows_week	dynamic_temporal_behavioral
Neural Discrete-Time Survival	week	discrete_time_index
Gradient-Boosted Weekly Hazard	studied_credits	discrete_time_index

Table 10 Calibration gap summary for Family A and Family B models. Values are IPCW-adjusted mean absolute gaps between predicted risk and observed event rate across quantile bins; lower values indicate better calibration.

Model	Calib@10	Calib@20	Calib@30
Family B			
Random Survival Forest	0.0121	0.0162	0.0138
DeepSurv	0.0320	0.0145	0.0119
DeepHit	0.0333	0.0322	0.0270
Weibull AFT	0.0369	0.0219	0.0123
Neural-MTLR	0.0384	0.0503	0.0521
Cox Comparable	0.0386	0.0183	0.0106
Gradient-Boosted Cox	0.0437	0.0246	0.0160
Royston-Parmar	0.0438	0.0319	0.0212
XGBoost AFT	0.1242	0.1775	0.2176
Family A			
Gradient-Boosted Weekly Hazard	0.0886	0.1122	0.1438
Linear Discrete-Time Hazard	0.0890	0.1123	0.1427
Neural Discrete-Time Survival	0.0915	0.1174	0.1490
CatBoost Weekly Hazard	0.0919	0.1161	0.1475
Poisson Piecewise-Exponential	0.0980	0.1023	0.1228

Declarations

Funding. This research received no external funding.

Ethical Approval and accordance. Not applicable. This study did not involve human research participants or live vertebrates. It is a secondary analysis of the publicly available, fully anonymized Open University Learning Analytics Dataset (OULAD), which contains only anonymized records and no identifiable personal data. No participants were recruited, no data were collected by the authors, and no intervention was administered. Accordingly, ethics committee/IRB approval was not required.

Consent to participate. Not applicable.

Consent to publish. Not applicable.

Data availability. The datasets analysed during the current study are available in the Open University Learning Analytics Dataset (OULAD) repository, https://research.stem.open.ac.uk/ouanalyse/dataset/. Direct download at http://schools.stem.open.ac.uk/cdn/files/anonymisedData.zip (Kuzilek et al., 2017).

Supplementary information. Online Resource 1 (`ESM_1.pdf`): hyperparameter search grids for all 14 models, including candidate count, selection criterion, and full grid specification. Online Resource 2 (`ESM_2_supplementary_appendix.pdf`): supplementary audits (evaluation protocol, calibration by horizon, preprocessing and tuning, bootstrap uncertainty and inferential scope, split/context overlap, discrete-time diagnostic bridge, proportional-hazards audit, early-window sensitivity) and reliability diagrams (Supplementary Tables S1–S10 and Figures S1–S4).

References

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A., Abid, M., ... Khan, S. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access*, *9*, 7519–7539, https://doi.org/10.1109/ACCESS.2021.3049446

Akçapınar, G., Altun, A., Aşkar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. *International Journal of Educational Technology in Higher Education*, *16*, , https://doi.org/10.1186/s41239-019-0172-z

Alalawi, K., Athauda, R., Chiong, R. (2024). An extended learning analytics framework integrating machine learning and pedagogical approaches for student performance prediction and intervention. *International Journal of Artificial Intelligence in Education*, *35*, 1239–1287, https://doi.org/10.1007/s40593-024-00429-7

Aljohani, N., Fayoumi, A., Hassan, S. (2019). Predicting at-risk students using clickstream data in the virtual learning environment. *Sustainability*, , https://doi.org/10.3390/su11247238

Allison, P.D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, *13*, 61–98, https://doi.org/10.2307/270718

Antolini, L., Boracchi, P., Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in Medicine*, *24*(24), 3927–3944, https://doi.org/10.1002/sim.2427

Arnold, K.F., Davies, V.A., De Kamps, M., Tennant, P.W.G., Mbotwa, J., Gilthorpe, M.S. (2020). Reflection on modern methods: Generalized linear models for prognosis and intervention—theory, practice and implications for machine learning. *International Journal of Epidemiology*, *49*, 2074–2082, https://doi.org/10.1093/ije/dyaa049

Bañeres, D., Rodríguez-González, M., Guerrero-Roldán, A.-E., Cortadas, P. (2023). An early warning system to identify and intervene online dropout learners. *International Journal of Educational Technology in Higher Education*, *20*, 1–25, https://doi.org/10.1186/s41239-022-00371-5

Burk, L., Zobolas, J., Bischl, B., Bender, A., Wright, M.N., Sonabend, R. (2024). A large-scale neutral comparison study of survival models on low-dimensional data. *arXiv preprint arXiv:2406.04098*, , https://doi.org/10.48550/arXiv.2406.04098

Cannistrà, M., Masci, C., Ieva, F., Agasisti, T., Paganoni, A. (2021). Early-predicting dropout of university students: An application of innovative multilevel machine learning and statistical techniques. *Studies in Higher Education*, , https://doi.org/10.1080/03075079.2021.2018415

Collins, G.S., Dhiman, P., Archer, L., Van Calster, B., Harrell, F.E., Martin, G.P., ... Riley, R.D. (2024). Evaluation of clinical prediction models (part 1): From development to external validation. *The BMJ*, *384*, e074819, https://doi.org/10.1136/bmj-2023-074819

Coussement, K., Phan, M., De Caigny, A., Benoit, D., Raes, A. (2020). Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. *Decision Support Systems*, *135*, 113325,

https://doi.org/10.1016/j.dss.2020.113325

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, *34*, 187–220, https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Dwivedi, V.P., Joshi, C.K., Laurent, T., Bengio, Y., Bresson, X. (2022). Benchmarking graph neural networks. *Journal of Machine Learning Research*, *23*, , Retrieved from https://jmlr.org/papers/volume24/22-0567/22-0567.pdf

Farhood, H., Joudah, I., Beheshti, A., Müller, S. (2024). Evaluating and enhancing artificial intelligence models for predicting student learning outcomes. *Informatics*, *11*(46), , https://doi.org/10.3390/informatics11030046

Fazil, M., & Albahlal, B.M. (2025). MultiFAR: Multidimensional information fusion with attention-driven representation learning for student performance prediction. *PLoS ONE*, *20*(10), e0333099, https://doi.org/10.1371/journal.pone.0333099

Fisher, A., Rudin, C., Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*, 1–81, Retrieved from https://jmlr.org/papers/v20/18-760.html

Gao, R., & Ni, Q. (2025). FairEduNet: A novel adversarial network for fairer educational dropout prediction. *Scientific Reports*, *15*, 34032, https://doi.org/10.1038/s41598-025-13632-w

Gerds, T.A., & Schumacher, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, *48*, 1029–1040, https://doi.org/10.1002/bimj.200610301

Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, *18*, 2529–2545, https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18⟨2529::AID-SIM274⟩3.0.CO;2-5

Gutierrez-Pachas, D.A., Garcia-Zanabria, G., Cuadros-Vargas, E., Camara-Chavez, G., Gomez-Nieto, E. (2023). Supporting decision-making process on higher education dropout by analyzing academic, socioeconomic, and equity factors through machine learning and survival analysis methods in the latin american context. *Education Sciences*, *13*(2), 154, https://doi.org/10.3390/educsci13020154

Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A. (1982). Evaluating the yield of medical tests. *JAMA*, *247*, 2543–2546, https://doi.org/10.1001/jama.1982.03320430047030

He, Y., Chen, R., Li, X., Hao, C., Liu, S., Zhang, G., Jiang, B. (2020). Online at-risk student identification using RNN-GRU joint neural networks. *Information*, *11*(10), 474, https://doi.org/10.3390/info11100474

Hooshyar, D., & Yang, Y. (2024). Problems with shap and lime in interpretable ai for education: A comparative study of post-hoc explanations and neural-symbolic rule extraction. *IEEE Access*, *12*, 137472–137490, https://doi.org/10.1109/ACCESS.2024.3463948

Ifenthaler, D., & Yau, J.Y.-K. (2020). Utilising learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development*, *68*, 1961–1990, https://doi.org/10.1007/s11423-020-09788-z

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9), 100804, https://doi.org/10.1016/j.patter.2023.100804

Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y. (2018). Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, *18*, , https://doi.org/10.1186/s12874-018-0482-1

Kim, S.Y. (2023). GNN-surv: Discrete-time survival prediction using graph neural networks. *Bioengineering*, *10*(9), 1046, https://doi.org/10.3390/bioengineering10091046

Kuzilek, J., Hlosta, M., Zdráhal, Z. (2017). Open university learning analytics dataset. *Scientific Data*, *4*, , https://doi.org/10.1038/sdata.2017.171

Kvamme, H., & Borgan, Ø. (2021). Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, , https://doi.org/10.1007/s10985-021-09532-6

Kvamme, H., Borgan, Ø., Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, *20*, 1–30, Retrieved from https://jmlr.org/papers/v20/18-424.html

Li, M., & Janson, L. (2024). Optimal ablation for interpretability. *arXiv*, , https://doi.org/10.5555/3737916.3741384 2409.09951

Lillelund, C.M., Qi, S.-a., Greiner, R., Pedersen, C.F. (2025, June). *Stop chasing the c-index: This is how we should evaluate our survival models.* Retrieved from https://arxiv.org/abs/2506.02075

Longato, E., Vettoretti, M., Di Camillo, B. (2020). A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. *Journal of Biomedical Informatics*, *108*, 103496, https://doi.org/10.1016/j.jbi.2020.103496

Marcolino, M., Porto, T., Primo, T., Targino, R., Ramos, V., Queiroga, E., ... Cechinel, C. (2025). Student dropout prediction through machine learning optimization: Insights from moodle log data. *Scientific Reports*, *15*, , https://doi.org/10.1038/s41598-025-93918-1

Mduma, N., Kalegele, K., Machuve, D. (2019). A survey of machine learning approaches and techniques for student dropout prediction. *Data Science Journal*, *18*, 14, https://doi.org/10.5334/dsj-2019-014

Mihăescu, M., & Popescu, P. (2021). Review on publicly available datasets for educational data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*, , https://doi.org/10.1002/widm.1403

Mubarak, A., Cao, H., Zhang, W. (2020). Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environments*, *30*, 1414–1433, https://doi.org/10.1080/10494820.2020.1727529

Nagy, M., & Molontay, R. (2023). Interpretable dropout prediction: Towards xai-based personalized intervention. *International Journal of Artificial Intelligence in Education*, , https://doi.org/10.1007/s40593-023-00331-8

Nooraldaim, A.S., Xue, Z., Lai, X., Zhu, X., Wang, J. (2026). Multi-task survival modeling of dependent failure and reimplantation events in dental implants. *Scientific Reports*, , https://doi.org/10.1038/s41598-026-40955-z (Advance online publication)

Oqaidi, K., Aouhassi, S., Mansouri, K. (2022). Towards a students' dropout prediction model in higher education institutions using machine learning algorithms. *International Journal of Emerging Technologies in Learning*, *17*, 103–117, https://doi.org/10.3991/ijet.v17i18.25567

Pan, F., Huang, B., Zhang, C., Zhu, X., Wu, Z., Zhang, M., ... Li, Z. (2022). A survival analysis based volatility and sparsity modeling network for student dropout prediction. *PLoS ONE*, *17*(5), e0267138, https://doi.org/10.1371/journal.pone.0267138

Park, S.Y., Park, J.E., Kim, H.J., Park, S.H. (2021). Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models. *Korean Journal of Radiology*, *22*, 1697–1707, https://doi.org/10.3348/kjr.2021.0223

Peng, R.D. (2011). Reproducible research in computational science. *Science*, *334*, 1226–1227, https://doi.org/10.1126/science.1213847

Prenkaj, B., Velardi, P., Stilo, G., Distante, D., Faralli, S. (2020). A survey of machine learning approaches for student dropout prediction in online courses. *ACM Computing Surveys*, *53*, , https://doi.org/10.1145/3388792

Rodríguez, P., Villanueva, A., Dombrovskaia, L., Valenzuela, J. (2023). A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: The case of chile. *Education and Information Technologies*, , https://doi.org/10.1007/s10639-022-11515-5

Sandoval-Palis, I., Naranjo, D., Vidal, J., Gilar-Corbí, R. (2020). Early dropout prediction model: A case study of university leveling course students. *Sustainability*, , https://doi.org/10.3390/su12229314

Seo, E., Yang, J., Lee, J., So, G. (2024). Predictive modelling of student dropout risk: Practical insights from a south korean distance university. *Heliyon*, *10*, , https://doi.org/10.1016/j.heliyon.2024.e30960

Sghir, N., Adadi, A., Lahmer, M. (2022). Recent advances in predictive learning analytics: A decade systematic review (2012–2022). *Education and Information Technologies*, , https://doi.org/10.1007/s10639-022-11536-0

Shiao, Y., Chen, C., Wu, K., Chen, B., Chou, Y., Wu, T. (2023). Reducing dropout rate through a deep learning model for sustainable education: Long-term tracking of learning outcomes of an undergraduate cohort from 2018 to 2021. *Smart Learning Environments*, *10*, , https://doi.org/10.1186/s40561-023-00274-6

Sønderlund, A.L., Hughes, E., Smith, J. (2018). The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology*, *50*, 2594–2618, https://doi.org/10.1111/bjet.12720

Suresh, K., Severn, C., Ghosh, D. (2022). Survival prediction models: An introduction to discrete-time modeling. *BMC Medical Research Methodology*, *22*, , https://doi.org/10.1186/s12874-022-01679-6

Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in finnish higher education. *Technology in Society*, , https://doi.org/10.1016/j.techsoc.2024.102474

Van Calster, B., McLernon, D.J., van Smeden, M., Wynants, L., Steyerberg, E.W. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, *17*(1), 230, https://doi.org/10.1186/s12916-019-1466-7

Vasilev, I., Petrovskiy, M., Mashechkin, I. (2023). Sensitivity of survival analysis metrics. *Mathematics*, *11*(20), 4246, https://doi.org/10.3390/math11204246

Villar, A., & de Andrade, C.R.V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. *Discover Artificial Intelligence*, *4*, 2, https://doi.org/10.1007/s44163-023-00079-z

Waheed, H., Hassan, S., Aljohani, N., Hardman, J., Alelyani, S., Nawaz, R. (2020). Predicting academic performance of students from vle big data using deep learning

models. *Computers in Human Behavior*, *104*, 106189, https://doi.org/10.1016/j.chb.2019.106189

Wiegrebe, S., Kopper, P., Sonabend, R., Bender, A. (2023). Deep learning for survival analysis: A review. *Artificial Intelligence Review*, , https://doi.org/10.1007/s10462-023-10681-3

Xing, W., & Du, D. (2019). Dropout prediction in moocs: Using deep learning for personalized intervention. *Journal of Educational Computing Research*, *57*, 547–570, https://doi.org/10.1177/0735633118757015

Zanellati, A., Zingaro, S.P., Gabbrielli, M. (2024). Balancing performance and explainability in academic dropout prediction. *IEEE Transactions on Learning Technologies*, *17*, 2140–2153, https://doi.org/10.1109/TLT.2024.3425959

Zhidkikh, D., Heilala, V., Van Petegem, C., et al. (2024). Reproducing predictive learning analytics in cs1: Toward generalizable and explainable models for enhancing student retention. *Journal of Learning Analytics*, *11*, 132–150, https://doi.org/10.18608/jla.2024.7979

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf
- ESM2supplementaryappendix.pdf