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 Distributed Data Fusion and Lightweight Deep Learning Framework for Hierarchical Risk Prediction over Heterogeneous Institutional Data Streams

Jinyun Hu
Dongying Vocational College
Xuenan Du
Dongying Vocational College
Wen Song
Dongying Vocational College
Jian Su
Dongying Vocational College
Na Li
Dongying Vocational College
Jiangbo Guo
Dongying Vocational College
Huayan Shang
Shyshy19810126.com

Dongying Vocational College

Research Article

Keywords: distributed data fusion, heterogeneous stream processing, LSTM-attention, hierarchical risk classification, lightweight inference

Posted Date: June 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9883305/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 Distributed Data Fusion and Lightweight Deep Learning Framework for Hierarchical Risk Prediction over Heterogeneous Institutional Data Streams

Jinyun Hu[1], Xuenan Du[2], Wen Song[2], Jian Su[3], Na Li[2], Jiangbo Guo[4], Huayan Shang[3*]

[1]School of Digital Finance and Management, Dongying Vocational College, Dongying, 257091, China.
[2]School of Cultural Tourism, Health Care and Education, Dongying Vocational College, Dongying, 257091, China.
[3]School of Public Education, Dongying Vocational College, Dongying, 257091, China.
[4]Dongying Vocational College, Dongying, 257091, China.

*Corresponding author(s). E-mail(s): Shyshy1981@126.com;
Contributing authors: hujinyun1216@126.com; 1032950096@qq.com; 845540559@qq.com; sujian328@163.com; 389859429@qq.com; jiangboguo79@163.com;

Abstract

Heterogeneous data streams produced by independently operated institutional systems impose a recurring bottleneck for scalable low-latency inference: schema heterogeneity, asynchronous updates, and access-control fragmentation often prevent direct end-to-end learning. We present a distributed multi-source data-fusion pipeline coupled with a lightweight Long Short-Term Memory network with additive attention (LSTM-Attention) for hierarchical four-level risk classification over such streams. The framework decomposes inference into five horizontally scalable phases: per-source extraction, stream merge, tensor construction, hierarchical labelling, and data-parallel training. Communication overhead is bounded by the parameter size during AllReduce, and the resulting 4.7 MB checkpoint supports deployment on commodity cluster and cloud infrastructure without

specialised hardware. On the institutional V100 baseline, the measured single-worker implementation attains an end-to-end throughput of 9,756 students/s. A complementary synthetic-cohort stress test exercises the distributed pipeline beyond the empirical cohort size and provides an engineering-reference characterisation of scale-out behaviour, reaching approximately 67,800 students/s under an 8-worker configuration with an estimated 86.9% parallel efficiency. As a vocational-education case study, a prospective longitudinal cohort of 4,000 students across four semesters (16,000 observations, 35 features, four source systems) achieves 92.3% accuracy, macro-F1 = **0.891**, and Cohen's $\boldsymbol{\kappa}$ = **0.890**, outperforming five baselines. Source-system ablation and SHAP attribution consistently identify industry-university integration features as contributing 46.6% of predictive power, while fault-tolerance evaluation shows graceful degradation under single-source outages, with accuracy loss bounded by 7.6 percentage points. Leakage-free external validation on OULAD ($\boldsymbol{n}$ = **32,593**) and EdNet-KT1 ($\boldsymbol{n}$ = **12,347**) yields quantified lower-bound performance under maximal distribution shift, with out-of-distribution accuracies of 0.595 and 0.601 and Cohen's $\boldsymbol{\kappa}$ values of 0.400 and 0.380 on the four-class problem.

Keywords: distributed data fusion, heterogeneous stream processing, LSTM-attention, hierarchical risk classification, lightweight inference

1 Introduction

1.1 Heterogeneous Multi-Source Data Streams in Modern Institutions

Distributed enterprise information systems present a well-documented data-engineering challenge: when multiple independently operated data stores — each with its own schema, update cadence, and access-control regime — must be fused for downstream analytical inference, the integration pipeline becomes the critical bottleneck rather than the model itself (Dean et al, 2012; Li et al, 2014). This challenge is particularly acute in modern educational institutions, which operate a portfolio of disconnected systems: academic management systems (AMS) for grades and credit accounting, learning management systems (LMS) for online interaction traces, student affairs systems (SAS) for demographic and attendance records, and — in vocational settings — industry-university cooperation platforms (IUCP) for workplace learning signals (Daniel, 2015; Ferguson, 2012). The result is a *heterogeneous, distributed data substrate* in which analytically relevant signals are scattered across systems that rarely communicate, and translating this multi-source landscape into a unified temporal representation for scalable deep-learning inference requires a principled pipeline architecture (Ferguson, 2012; Gašević et al, 2015). As institutional data volumes and source-system diversity grow, so do the importance of horizontal scalability, modular pipeline design, and lightweight inference suitable for cluster and cloud deployment (Zaharia et al, 2013).

This problem is particularly acute in vocational education under *industry-university integration* policies, where students alternate between theoretical campus instruction and practical enterprise-based training (Liu et al, 2023). The workplace-learning component of the curriculum generates a class of signals — internship hours, workplace supervisor ratings, skill-competition outcomes, vocational certification — that lives in the IUCP system and is largely absent from the LMS and AMS streams that drive most existing warning systems. Models that learn only from the classroom side of this divide systematically under-represent a category of at-risk students whose academic difficulty originates at the campus-workplace interface rather than in the classroom alone.

1.2 Academic Risk in Vocational Education

Academic difficulties are prevalent among students in Chinese vocational colleges, contributing to elevated dropout risk, delayed graduation, and weakened workforce–industry alignment (Liu et al, 2023). Prior research consistently documents the substantial cost of failing to identify struggling students early enough for effective intervention (Marbouti et al, 2016).

Traditional academic warning systems in vocational institutions rely predominantly on end-of-semester grade point averages (GPAs) and course-failure counts as triggering mechanisms (Jayaprakash et al, 2014). This reactive design suffers from temporal delay: by the time students receive warnings based on cumulative performance deficits, the opportunity for timely remediation has often passed. Moreover, conventional systems employ binary classification — warning vs. non-warning — that fails to capture the graduated spectrum of academic risk and provides insufficient guidance for differentiated intervention . The distinctive features of industry-university integration — enterprise internships, skill competitions, and vocational certifications — remain largely absent from these frameworks.

1.3 Deep Learning and Educational Data Mining

Recent advances in educational data mining and learning analytics have demonstrated the potential of machine learning approaches for predicting student academic outcomes (Romero and Ventura, 2020; Zhang and Aslan, 2024). Studies have applied logistic regression (Costa et al, 2017), random forests (Breiman, 2001; Márquez-Vera et al, 2016), gradient-boosted trees (Chen and Guestrin, 2016), and ensemble methods (Amrieh et al, 2016) to identify at-risk students. Recent contributions extend these techniques: Pecuchova and Drlik (2024) combine clustering with deep models for early dropout detection on VLE traces; Vaarma and Li (2024) report a large-scale empirical machine-learning dropout study in Finnish higher education; Marcolino et al (2025) and Mustofa et al (2025) apply optimised gradient-boosting and explainable AI to Moodle log streams; and Waheed et al (2024) benchmark deep models on VLE big-data. However, most of these methods consume static, cross-sectional feature vectors and fail to capture the temporal dynamics inherent in longitudinal student trajectories.

Deep learning architectures, particularly recurrent neural networks (RNNs) and their gated variants (Hochreiter and Schmidhuber, 1997; LeCun et al, 2015; Schmidhuber, 2015), offer markedly stronger capabilities for modelling sequential dependencies. LSTM networks have shown promise in MOOC dropout prediction (Fei and Yeung, 2015; Qiu et al, 2022), knowledge tracing (Piech et al, 2015), and learning-outcome forecasting (Wang et al, 2021). Bidirectional LSTM architectures (Graves and Schmidhuber, 2005) further enable models to integrate context from both past and (within the observation window) future time steps. Augmenting LSTM models with attention (Bahdanau et al, 2015; Vaswani et al, 2017) allows the network to learn data-driven weights over time steps, improving both accuracy and the interpretability of *which* point in a student's trajectory drove a given prediction. Heterogeneous and temporal extensions are emerging — Cheng et al (2024) model student trajectories as temporal heterogeneous information networks — yet the systematic fusion of independently operated institutional source systems under a horizontally scalable distributed pipeline, and the application of attention-augmented sequential deep learning to vocational education under industry-university integration, remain scarce (Yin et al, 2024).

1.4 Gaps Addressed and Contributions

The present study addresses four interrelated gaps. *First,* although a substantial literature on academic risk prediction exists in higher education, vocational settings — and in particular the industry-university integration context — are systematically underrepresented; workplace-learning signals are seldom incorporated into deployed warning systems. *Second,* most prior work employs cross-sectional or short-horizon designs that cannot model the cumulative dynamics of academic difficulty. *Third,* most existing systems are binary triage tools; they cannot calibrate the *intensity* of recommended intervention to the *severity* of risk. *Fourth,* very few studies report leakage-free external validation on independent benchmarks, leaving the question of how much reported accuracy reflects genuine signal versus over-fitting unanswered (Whitehill et al, 2015; Hellas et al, 2018).

This research develops and validates a deep-learning approach to early academic-risk identification in vocational education that addresses each of these gaps. The model is a bidirectional LSTM with additive attention that learns from longitudinal student trajectories spanning four consecutive semesters. It produces a four-level hierarchical classification — Normal, Blue (mild), Yellow (moderate), Red (severe) — explicitly designed for differentiated intervention rather than binary triage. The feature space fuses traditional academic and behavioural variables with workplace learning indicators sourced from the IUCP, providing a unified longitudinal representation. Model behaviour is decomposed via SHAP (Lundberg and Lee, 2017) to attribute predictive credit to feature categories, and the model is subjected to a leakage-free early-window external validation on two large public benchmarks (OULAD and EdNet-KT1).

The study makes the following contributions:

1. A **scalable multi-source data integration pipeline** that fuses four institutional source systems (AMS, LMS, SAS, IUCP) into a unified three-dimensional tensor

representation of each student's four-semester trajectory, suitable for end-to-end deep learning at institutional scale and engineered to allow new source systems to be added at the extractor level without re-engineering downstream model code.
2. A **hierarchical four-level warning scheme** with class-specific recall calibrated so that high-risk (Red) students are prioritised at the cost of mild over-flagging of borderline cases, supporting differentiated rather than binary intervention.
3. An **LSTM-Attention model for vocational risk prediction** that systematically incorporates industry-university integration variables alongside traditional academic indicators, with SHAP-based attribution quantifying the marginal contribution of workplace variables ($\approx 46.6\,\%$ of total predictive importance) — to our knowledge the first such decomposition reported for a Chinese vocational-education cohort.
4. A **leakage-free external validation protocol on two public benchmarks** (OULAD, $n = 32{,}593$; EdNet-KT1, $n = 12{,}347$ valid students) under strict early-window prediction (first $30\,\%$ of the OULAD course; first half of each EdNet interaction sequence). The external benchmarks characterise the *leakage-free, early-window, generic-feature lower bound* for this class of model.

Scope and positioning.

The reusable contribution of this paper is a distributed application-middleware layer for heterogeneous institutional streams — source-isolated extractors, partition-parallel merge, idempotent tensor construction, manifest-based fault tolerance, AllReduce-bounded data-parallel training, and lightweight inference on commodity infrastructure. The vocational-education cohort provides a high-dimensional, privacy-constrained, multi-source evaluation environment in which to characterise this layer empirically; the same pipeline abstraction is intended to apply to other institutional risk-inference settings (e.g., clinical deterioration monitoring, financial compliance scoring, network-anomaly triage) for which heterogeneous source systems and four-level severity stratification are the dominant engineering constraints. Section 4.5 reports two years of operational deployment experience under this framing.

The central research questions guiding this investigation are: **(RQ1)** Can a scalable LSTM-Attention framework that fuses heterogeneous multi-source educational data streams predict hierarchical four-level academic warning states more accurately than competitive baselines? **(RQ2)** What is the relative predictive contribution of industry-university integration variables compared with conventional academic and behavioural predictors? **(RQ3)** How early in a student's trajectory does the model achieve actionable predictive accuracy? **(RQ4)** To what extent do the architectural and behavioural-feature components of the framework generalise across markedly different educational contexts under a leakage-free protocol?

1.5 Related Work

Three lines of prior research are directly relevant to the present framework: distributed integration of heterogeneous multi-source data streams, sequential deep learning with attention mechanisms for temporal prediction, and hierarchical risk stratification with leakage-free evaluation.

1.5.1 Distributed Data Integration over Heterogeneous Streams

Integrating heterogeneous, independently operated data streams into a unified representation is a long-standing problem in distributed-systems and data-engineering research. Dean et al (2012) demonstrated that large-scale deep-learning pipelines require modular architectures in which data extraction and model computation are decoupled across nodes to avoid single-node bottlenecks. Li et al (2014) formalised this into the Parameter Server paradigm — an extractor-aggregator-model tripartite decomposition — that we adapt as the per-source extractor design of our integration pipeline. For streaming contexts, Zaharia et al (2013) showed that fault-tolerant analytics over heterogeneous sources requires a discretised micro-batch ingestion layer that isolates source-level variability from downstream model training. In the specific context of institutional data systems, Daniel (2015) and Ferguson (2012) characterise the "data silo" problem in which analytically relevant signals are fragmented across independently maintained operational databases with incompatible schemas and update cadences; Gašević et al (2015) document the data-engineering bottleneck as the principal impediment to deployable learning analytics at scale. The present pipeline advances this line of work by providing a horizontally scalable, modular architecture explicitly designed for semester-level temporal tensor construction, exposing a stable interface contract that enables source-system replacement without downstream re-engineering. Recent work on Spark-based educational data stream processing (Li et al, 2024) and on scalability challenges in multimodal learning analytics (Yan et al, 2024) further motivates the engineering choices made here. Complementary work in distributed dataflow processing (Akidau et al, 2015) and parameter-server communication (Li et al, 2014) establishes the broader infrastructure layer on which such pipelines depend; the present contribution occupies the application middleware layer between those infrastructure primitives and the institutional source systems.

1.5.2 Sequential Deep Learning and Attention Mechanisms

LSTM networks (Hochreiter and Schmidhuber, 1997) and their GRU variant (Cho et al, 2014) are the established architectures for modelling longitudinal data with long-range temporal dependencies, with bidirectional extensions (Graves and Schmidhuber, 2005) providing context integration from both preceding and following time steps. LeCun et al (2015) and Schmidhuber (2015) establish the theoretical foundations for deep sequential learning. Additive-attention mechanisms (Bahdanau et al, 2015), later generalised in the Transformer architecture (Vaswani et al, 2017), augment recurrent encoders with data-driven time-step importance weights that improve both predictive accuracy and temporal interpretability. Within educational applications of sequential deep learning, Fei and Yeung (2015) pioneered LSTM-based MOOC dropout prediction, Piech et al (2015) introduced deep knowledge tracing, Wang et al (2021) applied recurrent models to completion forecasting, and Qiu et al (2022) demonstrated process-behaviour predictive analytics on e-learning platforms. These prior applications, however, consume single-source LMS streams and do not address the multi-system integration challenge that the present pipeline solves. The application of attention-augmented sequential architectures to multi-source heterogeneous streams in

resource-constrained deployment settings remains an active open problem, motivating the lightweight design and evaluation protocol presented here.

1.5.3 Hierarchical Risk Stratification and Leakage-Free Evaluation

Binary at-risk / not-at-risk classification has been the dominant output formulation in educational early-warning research (Hellas et al, 2018; Whitehill et al, 2015), yet binary outputs are insufficient for systems that must allocate differentiated interventions calibrated to severity (Arnold and Pistilli, 2012). Graduated stratification has been explored in academic-warning policy literature (Pistolesi, 2017) and in clinical deterioration monitoring, but the combination of a learnt deep model and a four-tier output for institutional academic risk remains under-explored. A more pressing methodological gap concerns evaluation rigour: Hellas et al (2018) and Whitehill et al (2015) document the pervasive contamination of reported accuracy by post-outcome signals — such as end-of-course click totals included as predictors of course outcomes — that constitute target leakage. Foundational educational data mining surveys (Amrieh et al, 2016; Márquez-Vera et al, 2016; Romero and Ventura, 2020) report predominantly internal-only validation results, with rare cross-dataset benchmarks under strict early-window cut-offs. The present work addresses all three limitations simultaneously: graduated four-level stratification, multi-source stream fusion, and leakage-free external validation under strict early-window cut-offs on two large-scale public benchmarks.

The remainder of this paper is structured as follows. Section 2 details the methodology, including the multi-source data integration pipeline, variable specification, model architecture, ablation protocol, statistical testing strategy, and external validation protocol. Section 3 presents empirical results across the institutional cohort and the two external benchmarks. Section 4 discusses theoretical implications, cross-context generalisation, practical scalability of the framework, deployment experience, and limitations. Section 5 provides concluding remarks, policy recommendations, and directions for future work.

2 Materials and Methods

2.1 Study Design and Ethical Considerations

This study employed a prospective longitudinal cohort design to track the academic trajectories of vocational college students over four consecutive semesters from September 2021 to June 2023. The research was conducted in accordance with institutional ethical guidelines and was reviewed and authorised by the Academic Affairs Office of Dongying Vocational College, which served as the institutional data governance authority. All student records were de-identified prior to analysis, with unique alphanumeric codes replacing personal identifiers. Given the retrospective character of the administrative data, individual informed consent requirements were waived under institutional policy for secondary analysis of anonymised administrative educational data (Pardo and Siemens, 2014). The research posed no greater than minimal risk to participants and was conducted solely for educational research purposes.

Figure 1 illustrates the overall framework: four heterogeneous source systems feed parallel extractors that emit columnar shards; a distributed stream merger partitions the join on (student_id, semester); the tensor constructor produces $\mathbf{X} \in \mathbb{R}^{N \times T \times F}$ for the BiLSTM-Attention inference engine that emits four hierarchical warning levels.

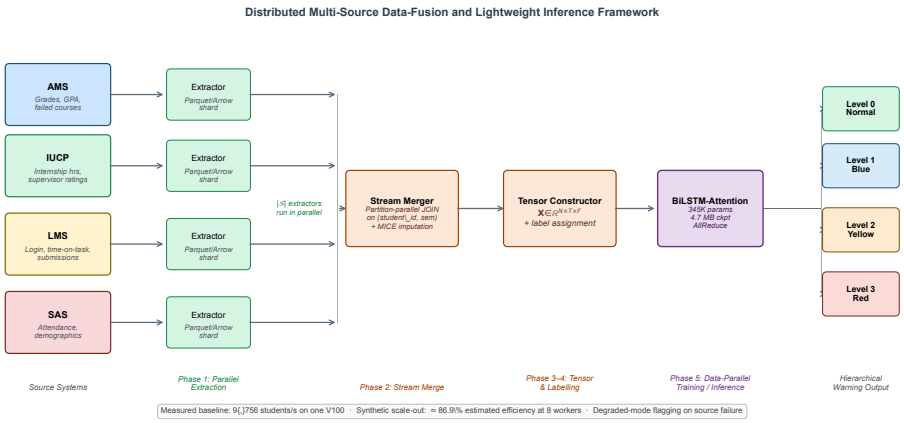


Fig. 1 Top-level architecture of the distributed multi-source data-fusion and lightweight inference framework. Source systems (AMS, IUCP, LMS, SAS) on the left feed parallel Phase-1 extractors, whose Parquet/Arrow shards flow into a partition-parallel stream merger and tensor constructor, terminating in a BiLSTM-Attention inference engine ($\approx 345\,\mathrm{K}$ parameters, 4.7 MB checkpoint) that emits a four-level hierarchical warning. The measured single-worker baseline is 9,756 students/s on one V100; a complementary synthetic-cohort stress test provides an engineering reference for multi-worker scale-out behaviour (Section 3.9.1).

2.2 Participants and Setting

The study sample comprised 4,000 students enrolled in the 2020 academic cohort (n = 2,000) and the 2021 academic cohort (n = 2,000) at a provincial-level demonstration vocational technical college in central China. Participants were distributed across eight major programmes: Mechanical Manufacturing (14.2 %), Electronic Commerce (13.8 %), Computer Applications (12.1 %), Construction Engineering (11.5 %), Tourism Management (12.6 %), Accounting (13.3 %), Automotive Maintenance (11.7 %), and Electrical Automation (10.8 %). Gender distribution was 52 % male and 48 % female. Mean age at enrolment was 18.7 years ($SD = 1.2$).

Inclusion criteria required full-time enrolment status throughout the observation period, participation in college-enterprise cooperative education programmes, completion of at least 70 % of scheduled coursework per semester, and availability of complete academic records. Application of these criteria resulted in the exclusion of 147 students (3.5 % of the sampling frame), yielding the final analytical sample of 4,000 students contributing 16,000 semester-level observations.

2.3 Distributed Multi-Source Data Integration Pipeline

Data were extracted from four integrated institutional information systems.

Academic Management System (AMS). Semester-wise records of course enrolments, examination scores, grade assignments, credit accumulation, and degree-audit status.

Industry-University Cooperation Platform (IUCP). Enterprise-based learning activities: internship placement records, duration logs, enterprise supervisor evaluations, project participation indicators, and industry-recognised skill assessment outcomes.

Learning Management System (LMS). Digital learning behaviours: login frequencies, time-on-task metrics, assignment submission records, discussion forum participation, and resource access patterns.

Student Affairs System (SAS). Demographic characteristics, socioeconomic background indicators, residential information, campus organisation involvement, and attendance monitoring data.

Data extraction occurred at four temporal milestones aligned with semester-conclusion dates: T1 (December 2021), T2 (June 2022), T3 (December 2022), and T4 (June 2023). The integration pipeline was deliberately modular: each source system was wrapped by an extractor that exposed a tabular semester-level view, and the four views were joined on (`student_id`, `semester`) into a unified semester-level table that fed the downstream tensor construction stage. This modular layout ensures that adding a fifth source — or replacing an existing source after a system upgrade — does not require re-engineering the downstream model.

Distributed extraction and horizontal scalability. Each source-system extractor operates as an independent process, enabling concurrent fan-out retrieval from distributed campus servers. The normalised semester-level views are materialised in a column-oriented format amenable to parallel read by distributed computation frameworks — such as Apache Spark — at institutions where student population or data volume exceeds the single-node processing ceiling (Zaharia et al, 2013). Horizontal scaling to a new source requires only the addition of a new extractor module conforming to the `(student_id, semester, feature_vector)` interface contract; the tensor constructor, model training, and inference stages require no modification. This decoupled design follows established distributed software-tool engineering principles (Dean et al, 2012; Li et al, 2014).

Algorithm 1 summarises the end-to-end distributed procedure.

Asymptotic Complexity and Communication Cost

Let N denote the cohort size, $T = |\mathcal{T}|$ the number of temporal cut-offs, F the total feature dimension, H the LSTM hidden size, $|\mathcal{S}|$ the number of source systems, and P the number of compute workers. Table 1 reports per-phase work, depth (critical path under ideal parallel scheduling), and inter-worker communication cost.

Three properties of this complexity profile are relevant to cluster-scale deployment. *First,* the dominant work term scales linearly in N for Phases 1–4 and in N/P for the

Algorithm 1 Distributed Multi-Source Stream Fusion Pipeline

Require: $\mathcal{S} = \{\text{AMS, IUCP, LMS, SAS}\}$ ▷ source-system endpoints
Require: $\mathcal{T} = \{t_1, t_2, t_3, t_4\}$ ▷ semester cut-off dates
Require: $\mathcal{U} = \{u_1, \ldots, u_N\}$ ▷ enrolled student ids
Ensure: $\mathbf{X} \in \mathbb{R}^{N \times |\mathcal{T}| \times F}$ ▷ student-time-feature tensor
Ensure: $\mathbf{Y} \in \{0, 1, 2, 3\}^{N \times |\mathcal{T}|}$ ▷ hierarchical risk labels
Ensure: $\hat{\theta}$ ▷ trained LSTM-Attention weights

Phase 1 — Parallel Extraction (one process per source)
for $s \in \mathcal{S}$ **in parallel do**
  $V_s \leftarrow \textsc{Extractor}_s(\mathcal{U}, \mathcal{T})$ ▷ tabular view, key=(u, t)
  $V_s \leftarrow \textsc{Normalise}(V_s)$ ▷ unit/scale harmonisation
  $\textsc{Persist}(V_s)$ to columnar store ▷ Parquet / Arrow shard
end for
Phase 2 — Stream Merge (parallel by student partition)
for partition p of $\mathcal{U}$ **in parallel do**
  $M_p \leftarrow \text{Join}_{(u,t)}(V_{\text{AMS}}, V_{\text{IUCP}}, V_{\text{LMS}}, V_{\text{SAS}})$ on partition p
  $M_p \leftarrow \textsc{Impute}(M_p)$ ▷ MICE for IUCP, median elsewhere
end for
$M \leftarrow \bigcup_p M_p$
Phase 3 — Tensor Construction
for $u \in \mathcal{U}$, $t \in \mathcal{T}$ **do**
  $\mathbf{X}[u, t, :] \leftarrow \textsc{FeatureVector}(M, u, t)$
end for
$\mathbf{X} \leftarrow \textsc{Standardise}(\mathbf{X})$ ▷ training-set mean/SD only
Phase 4 — Hierarchical Label Assignment (parallel by t)
for $t \in \mathcal{T}$ **in parallel do**
  for $u \in \mathcal{U}$ **do**
    $\mathbf{Y}[u, t] \leftarrow \textsc{HierarchicalRule}(\text{GPA}, \text{fails}, \text{attendance})$
  end for
end for
Phase 5 — Model Training (single GPU or distributed)
$\hat{\theta} \leftarrow \arg\min_\theta \ \mathcal{L}_{\text{CE-weighted}}(f_\theta(\mathbf{X}_{\text{train}}), \mathbf{Y}_{\text{train}})$
return $\mathbf{X}, \mathbf{Y}, \hat{\theta}$

data-parallel training stage, so doubling cohort size doubles total work but leaves per-worker wall-clock invariant under matching P. *Second,* inter-worker communication is bounded by either (i) a one-time shuffle of $\mathcal{O}(NTF/P)$ in Phase 2 or (ii) $\mathcal{O}(|\theta|)$ AllReduce traffic per training step, both of which are small relative to per-worker compute and therefore non-bottlenecking for institutional scale ($N \leq 10^5$). *Third,* the depth of Phases 1–4 is independent of $|\mathcal{S}|$ under per-source parallelism, so adding a fifth or sixth source-system increases total work but not critical-path latency. These three properties collectively justify the design claim that horizontal scale-out is achieved by adding extractors and workers without re-engineering the downstream stages.

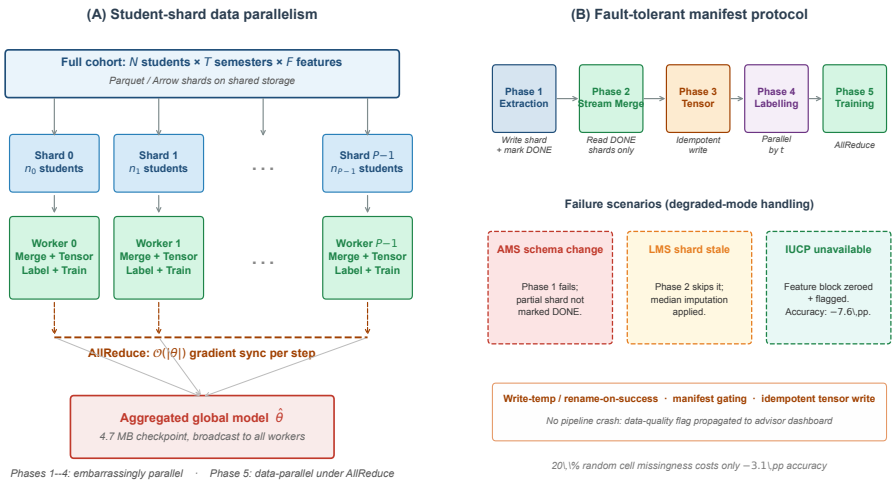


Fig. 2 Distributed execution model for Algorithm 1. Panel (A): student-shard data parallelism — the cohort is partitioned across P workers, each of which independently runs Phases 1–4 with no cross-worker communication; Phase 5 uses AllReduce to synchronise gradients of $\mathcal{O}(|\theta|) \approx 345\,\text{K}$ parameters per step, after which the aggregated 4.7 MB checkpoint is broadcast. Panel (B): fault-tolerant manifest protocol — each Phase-1 extractor writes a temporary shard and atomically renames it on success, Phase-2 reads only shards marked DONE in the manifest, Phase-3 tensor writes are idempotent. Three representative failure modes (AMS schema change, stale LMS shard, IUCP unavailable) all result in flagged degraded-mode predictions rather than pipeline crashes.

Table 1 Per-Phase Asymptotic Complexity of the Distributed Pipeline

Phase	Work	Depth (parallel)	Communication
1. Extraction (per-source)	$\mathcal{O}(NTF)$	$\mathcal{O}(NTF/\lvert\mathcal{S}\rvert)$	0 (local shards)
2. Stream merge (partitioned)	$\mathcal{O}(NT\lvert\mathcal{S}\rvert \log N)$	$\mathcal{O}((N/P)\,T\,\lvert\mathcal{S}\rvert \log(N/P))$	$\mathcal{O}(NTF/P)$ shuffle
3. Tensor construction	$\mathcal{O}(NTF)$	$\mathcal{O}(NTF/P)$	0 (data-local)
4. Hierarchical labelling	$\mathcal{O}(NT)$	$\mathcal{O}(NT/P)$	0
5. Training (data-parallel)	$\mathcal{O}(N(TFH + T^2H))$	$\mathcal{O}((N/P)(TFH + T^2H))$	$\mathcal{O}(\lvert\theta\rvert)$ AllReduce

$|\theta| \approx 345\,\text{K}$ for the proposed model; $T = 4$ semesters renders the T^2 attention term negligible. Phases 1–4 are embarrassingly parallel along the student axis; Phase 5 communication is bounded by parameter count, not data volume.

2.4 Variable Specification and Operationalisation

The analytical framework incorporated 35 predictor variables organised into six conceptual domains, plus two outcome variables. Table 2 provides variable definitions, measurement specifications, and distributional properties.

Table 2: Variable Definitions, Measurement Scales, and Descriptive Statistics

Category	Variable	Scale/Range	M	SD
Demographics	Gender	0=F, 1=M	0.52	0.50
	Age	17–22 yrs	18.70	1.21
	Major Category	1–8	—	—
	Geographic Origin	0=Urban, 1=Rural	0.65	0.48
	Family Economic Status	1–4 ordinal	2.18	0.82
	Parental Education	1–4 ordinal	2.05	0.91
	Entrance Score	Normalised 0–1	0.60	0.15
	Student Leadership	0/1	0.10	0.30
Academic Performance	Semester GPA	0.0–4.0	2.62	0.51
	Theory Course Average	0–100	86.36	8.83
	Practice Course Average	0–100	91.62	9.18
	Failed Courses Count	0–6	0.82	1.15
	Required Course Pass Rate	0.0–1.0	0.89	0.15
	Elective Credits Completed	0–10	6.15	2.03
Industry-University	Enterprise Internship Hours	0–160/sem	57.71	45.50
	Enterprise Supervisor Rating*	0–100	71.25	34.82
	Skill Competition Participation	0–5/sem	0.68	0.95
	Skill Competition Awards	0/1	0.24	0.43
	Vocational Certificates Earned	0–3 cumul.	1.12	0.98
	Coop Course Grade*	0–100	84.33	11.26
	Enterprise Project Involvement	0/1	0.35	0.48
Learning Behaviour	Class Attendance Rate	0.5–1.0	0.731	0.134
	Online Learning Hours	0–200/sem	58.42	28.15
	Assignment Submission Rate	0.0–1.0	0.812	0.152
	Library Visits	0–20/sem	6.23	4.58
	Classroom Interaction Freq.	0–50/sem	14.85	9.72
	LMS Login Days	20–120/sem	76.54	22.38
Extracurricular	Student Organisation Count	0–3	0.88	0.79

Category	Variable	Scale/Range	M	SD
	Volunteer Service Hours	0–40/sem	10.26	8.93
Temporal Features	Previous Semester GPA	0.0–4.0	2.65	0.53
	GPA Change Rate	−1.0 to 1.0	−0.012	0.185
	Grade Volatility	0.0–0.5	0.058	0.042

Note. $N = 16{,}000$ semester observations from 4,000 students. Variables marked with * are computed only for students with non-zero values; overall means include zeros for non-participants.

2.4.1 Outcome Variable: Hierarchical Warning Classification

Academic warning status was operationalised through a four-level ordinal classification system informed by established institutional policies and educational literature on academic risk stratification (Pistolesi, 2017; Stinebrickner and Stinebrickner, 2014). Each semester, students were assigned to one of four warning categories:

- **Level 0 (Normal Status).** GPA ≥ 2.5, zero failed courses, and attendance rate ≥ 0.80.
- **Level 1 (Blue Warning — Mild Difficulty).** $2.0 \leq$ GPA < 2.5 OR failed courses ≥ 2 OR ($2.3 \leq$ GPA < 2.5 AND failed courses $= 1$).
- **Level 2 (Yellow Warning — Moderate Difficulty).** $1.5 \leq$ GPA < 2.0 AND failed courses ≥ 2; OR (GPA < 2.0 AND failed courses ≥ 3); OR attendance rate < 0.65.
- **Level 3 (Red Warning — Severe Difficulty).** GPA < 1.5 OR failed courses ≥ 5 OR attendance rate < 0.55 OR composite risk score ≥ 7.

The distribution of warning classifications across the four semesters appears in Table 3.

Table 3 Distribution of Hierarchical Warning Classifications Across Semesters

Warning Level	Sem. 1 n (%)	Sem. 2 n (%)	Sem. 3 n (%)	Sem. 4 n (%)	Total n (%)
Level 0 (Normal)	2,711 (67.8)	2,691 (67.3)	2,653 (66.3)	2,570 (64.3)	10,625 (66.4)
Level 1 (Blue)	701 (17.5)	658 (16.5)	711 (17.8)	717 (17.9)	2,787 (17.4)
Level 2 (Yellow)	126 (3.2)	128 (3.2)	136 (3.4)	166 (4.2)	556 (3.5)
Level 3 (Red)	462 (11.6)	523 (13.1)	500 (12.5)	547 (13.7)	2,032 (12.7)
Total	4,000 (100)	4,000 (100)	4,000 (100)	4,000 (100)	16,000 (100)

$\chi^2(9) = 22.57$, $p = .007$; red-warning prevalence increased from 11.6 % (Semester 1) to 13.7 % (Semester 4).

2.4.2 Industry-University Integration Variables

A distinguishing feature of this study is the incorporation of enterprise-based learning metrics unique to vocational education under industry-university integration policies (Section 1.1). Enterprise internship participation increased progressively across semesters: 40 % of students completed internships in Semester 1, rising to 85 % by Semester 4. Enterprise supervisor ratings averaged 76.86 ($SD = 8.44$) on a 100-point scale among intern participants.

2.4.3 Temporal Feature Engineering

Three temporal features were engineered to capture longitudinal dynamics. **Previous Semester GPA** (GPA_{t-1}): the most proximal predictor of current academic standing. For Semester 1, entrance examination scores were transformed to the 4.0 scale using $\text{GPA}_{\text{initial}} = 0.5 + 3.5 \times \text{Entrance_Score_normalised}$. **GPA Change Rate**: $\Delta\text{GPA}_t = (\text{GPA}_t - \text{GPA}_{t-1})/(\text{GPA}_{t-1} + 0.01)$. **Grade Volatility**: $V_t = \text{SD}(\text{Course_Grades}_t)/(\text{Mean}(\text{Course_Grades}_t) + 0.01)$.

2.5 Distributed Fusion Architecture and Lightweight Prediction Model

2.5.1 Data Preprocessing Pipeline

Missing Value Imputation. Variables with structural missingness (enterprise supervisor rating, cooperative course grade) were addressed through conditional imputation. Cooperative course grades were imputed using MICE with 10 iterations and 5 imputation sets (van Buuren and Groothuis-Oudshoorn, 2011). Other missing values ($< 2\,\%$ of observations) were imputed using the median (continuous) or mode (categorical).

Feature Scaling. All continuous variables were Z-score standardised using training-set parameters only, to prevent data leakage into validation and test sets.

Temporal Sequence Construction. The resulting data structure took the form of a three-dimensional tensor of shape $(N_{\text{students}}, T_{\text{semesters}}, F_{\text{features}}) = (4000, 4, 35)$.

Class Imbalance Handling. Inverse-frequency class weights were incorporated in the loss function (He and Garcia, 2009):

$$w_c = \frac{N_{\text{total}}}{C \times N_c}, \tag{1}$$

yielding weights $\{w_0 = 0.60, w_1 = 1.43, w_2 = 7.19, w_3 = 1.97\}$ for the four warning levels.

Data Partitioning. Students (not observations) were randomly allocated to training (70 %, $n = 2{,}800$), validation (15 %, $n = 600$), and test (15 %, $n = 600$) sets using stratified sampling.

2.5.2 LSTM-Attention Model Specification

The prediction model employed a bidirectional LSTM augmented with an additive attention mechanism (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005; Bahdanau et al, 2015). The architecture comprised:

1. **Input Layer**: accepts tensor of shape $(N_{\text{batch}}, 4, 35)$.
2. **BiLSTM Layer 1**: 128 hidden units, return sequences; output shape $(N_{\text{batch}}, 4, 256)$.
3. **Dropout Layer 1**: rate 0.3 (Srivastava et al, 2014).
4. **BiLSTM Layer 2**: 64 hidden units, return sequences; output shape $(N_{\text{batch}}, 4, 128)$.
5. **Attention Mechanism**: let $\mathbf{H} = [h_1, h_2, h_3, h_4]$ denote the LSTM hidden states. Attention weights α_t are computed as

$$e_t = \mathbf{v}^\top \tanh(\mathbf{W}h_t + \mathbf{b}); \quad \alpha_t = \frac{\exp(e_t)}{\sum_j \exp(e_j)}, \tag{2}$$

   and the context vector is $\mathbf{c} = \sum_t \alpha_t h_t$.
6. **Dense Layer**: 32 neurons, ReLU activation.
7. **Dropout Layer 2**: rate 0.3.
8. **Output Layer**: 4 neurons, softmax activation.

2.5.3 Model Training Configuration

Loss function: categorical cross-entropy with class weights $\mathcal{L} = -(1/N)\sum_i w_{y_i} \sum_k y_{ik} \log \hat{y}_{ik}$. **Optimiser**: Adam (Kingma and Ba, 2015) with $\eta = 0.001$. **Batch size**: 64. **Maximum epochs**: 100; early stopping patience 15. **Regularisation**: L2 penalty $\lambda = 0.001$.

2.5.4 Baseline Models

Five baseline models were developed using identical preprocessing: (1) **Logistic Regression** (multinomial, L2, $C = 1.0$); (2) **Random Forest** (Breiman, 2001) (200 trees, max depth 15); (3) **XGBoost** (Chen and Guestrin, 2016) (200 estimators, lr = 0.1, max depth 6); (4) **Vanilla LSTM** (two-layer BiLSTM 128/64 without attention); (5) **GRU** (Cho et al, 2014) (matching the LSTM configuration).

2.5.5 Ablation Study Design

The comparison panel was designed as an ablation cascade. Removing the attention head from the LSTM-Attention model yields the Vanilla LSTM row, isolating the attention contribution. Replacing LSTM cells with GRU cells isolates the LSTM-vs-GRU contribution. Removing the recurrent architecture entirely (XGBoost, Random Forest, Logistic Regression) isolates the contribution of temporal modelling. The early-warning analysis (Section 2.6.3) constitutes a temporal-window ablation, and external validation (Section 2.7) constitutes a feature-set ablation.

2.5.6 Computational Complexity and Model Footprint

The inference model is intentionally compact. Parameter counts decompose as follows. For a bidirectional LSTM with input dimension d_{in} and hidden dimension d_h, the parameter count is $2 \times 4 \times (d_{\text{in}} + d_h + 1) \times d_h$. With $d_{\text{in}} = 35$, $d_h = 128$: BiLSTM Layer 1 contributes **167,936** parameters. With $d_{\text{in}} = 256$, $d_h = 64$: BiLSTM Layer 2 contributes **164,352** parameters. The additive-attention block ($\mathbf{W} \in \mathbb{R}^{128\times64}$, $\mathbf{v} \in \mathbb{R}^{64}$, $\mathbf{b} \in \mathbb{R}^{64}$) contributes **8,320** parameters. The two-layer MLP head ($128 \to 32 \to 4$) contributes **4,228** parameters. The **total is** $\approx$ 344,836 $\approx$ 345 **K parameters**, corresponding to a 32-bit float footprint of ≈ 1.38 MB.

2.6 Model Evaluation

2.6.1 Performance Metrics

Multi-class metrics were used: overall accuracy; macro-averaged precision, recall, and F1; class-specific precision, recall, and F1; one-vs-rest AUC-ROC; and Cohen's κ (Cohen, 1960) interpreted following Landis and Koch (1977).

2.6.2 Feature Importance Analysis

SHAP values were computed (Lundberg and Lee, 2017) on a random sample of 500 test-set predictions. Global feature importance was defined as the mean absolute SHAP value across instances.

2.6.3 Early Warning Capability Assessment

Three temporal scenarios were analysed: (1) Semester 1 data $\to$ Semester 2 status; (2) Semesters 1–2 $\to$ Semester 3 status; (3) Semesters 1–3 $\to$ Semester 4 status.

2.6.4 Statistical Significance Testing

Pairwise comparisons used McNemar's test on binary correct/incorrect outcomes across the shared test set (Dietterich, 1998). Bonferroni-corrected family-wise threshold: $\alpha_{\text{corrected}} = 0.01$ ($k = 5$ comparisons). Bootstrap confidence intervals (1,000 resamples, percentile method) accompanied the headline accuracy and macro-F1 estimates.

2.7 External Validation Protocol

External validation was conducted on the **Open University Learning Analytics Dataset (OULAD)** (Kuzilek et al, 2017) ($n = 32{,}593$) and **EdNet-KT1** (Choi et al, 2020) ($n = 12{,}347$ valid students). Both datasets lack industry-university integration signals, constituting a stringent out-of-distribution test.

Leakage-free early-window protocol.

- **OULAD**: Features derived only from days 0–80 (first $\approx 30\,\%$ of a 270-day presentation). Risk label mapped from `final_result`: Distinction $\to$ 0; Pass $\to$ 1; Fail $\to$ 2; Withdrawn $\to$ 3. Predictors: 7 demographic + 4 early-VLE + 2 early-assessment = 13 features.

- **EdNet-KT1**: Features from the first half of each student's interaction sequence. Risk label from second-half mean correctness ($\geq 0.85 \rightarrow 0$; $\geq 0.70 \rightarrow 1$; $\geq 0.50 \rightarrow 2$; $< 0.50 \rightarrow 3$). Students with < 20 interactions excluded.

For both datasets: 70/30 stratified train/test split (seed = 42); 4–5 models trained with identical hyperparameters; PyTorch-implemented streamlined LSTM-Attention (single BiLSTM-64 + attention + MLP head) trained for up to 40 epochs with early stopping (patience = 8).

2.8 Implementation, Reproducibility, and Fault Tolerance

Software stack. Python 3.9; Pandas 1.5; optional Apache Spark 3.3; TensorFlow 2.10 (Keras API); PyTorch 1.13 (external validation); SHAP v0.41; scikit-learn 1.1 (Pedregosa et al, 2011). Hardware: Ubuntu 20.04 LTS, NVIDIA Tesla V100 (32 GB VRAM, CUDA 11.7, cuDNN 8.5).

Complete hyperparameter specification is provided in Table 4. Random seed 42 was fixed for Python, NumPy, TensorFlow, and PyTorch prior to any partition or initialisation step.

Table 4 Complete Hyperparameter Specification for the LSTM-Attention Model

Hyperparameter	Value	Selection Method
BiLSTM Layer 1 hidden units	128	Grid search ({64, 96, 128, 192})
BiLSTM Layer 2 hidden units	64	Grid search ({32, 64, 96})
Dropout rate (both layers)	0.3	Grid search ({0.2, 0.3, 0.4})
Attention key dimension	64	Set equal to layer-2 hidden size
Dense head neurons	32	Grid search ({16, 32, 64})
Learning rate	0.001	Grid search ($\{3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}\}$)
L2 penalty λ	0.001	Grid search ($\{10^{-4}, 10^{-3}, 10^{-2}\}$)
Batch size	64	Fixed
Maximum epochs	100	Fixed
Early-stopping patience	15	Fixed
Random seed	42	Fixed across all experiments
Class weights	Inverse-frequency	Eq. (1)

Fault tolerance. Each Phase-1 extractor writes its normalised Parquet shard atomically (write-to-temp, rename-on-success). Phase 2 reads only shards marked complete by a manifest file, so a partial extraction results in a reduced-feature tensor rather than a pipeline crash. Tensor construction is idempotent: re-running Phase 3 overwrites the tensor file only if all four source shards are present and consistent. At inference time, if a source shard is stale, the corresponding feature block is imputed with the training-set column median and the warning output is annotated with a data-quality flag for advisor review. This degraded-mode behaviour ensures that a

temporary failure of a single institutional system does not cascade into a system-wide prediction outage.

Privacy and data minimisation. The extraction layer applies de-identification at the point of export: student names, national ID numbers, and contact details are replaced by opaque hash keys before any record is written to shared storage. An optional federated-learning extension (McMahan et al, 2017) is a natural next step for multi-institution deployments.

3 Results

3.1 Descriptive Statistics and Sample Characteristics

The final analytical sample comprised 4,000 vocational college students contributing 16,000 semester-level observations. Mean semester GPA was 2.62 ($SD = 0.51$). Enterprise internship participation increased from 39.3 % (Semester 1, mean 9.8 h) to 100 % (Semester 4, mean 109.6 h). Skill competition participation occurred in 18 % of semester observations; 5 % resulted in awards. Vocational certificate attainment grew cumulatively from $M = 0.25$ in Semester 1 to $M = 1.20$ in Semester 4.

3.2 Hierarchical Warning Classification Distribution

Table 3 (Section 2.4) summarises warning-level distributions. Red-warning prevalence increased from 11.6 % to 13.7 % across semesters ($\chi^2(9) = 22.57$, $p = .007$), consistent with cumulative-difficulty patterns in vocational education (Sneyers and De Witte, 2018). Figure 3 visualises the stacked-bar distribution and at-risk prevalence trends.

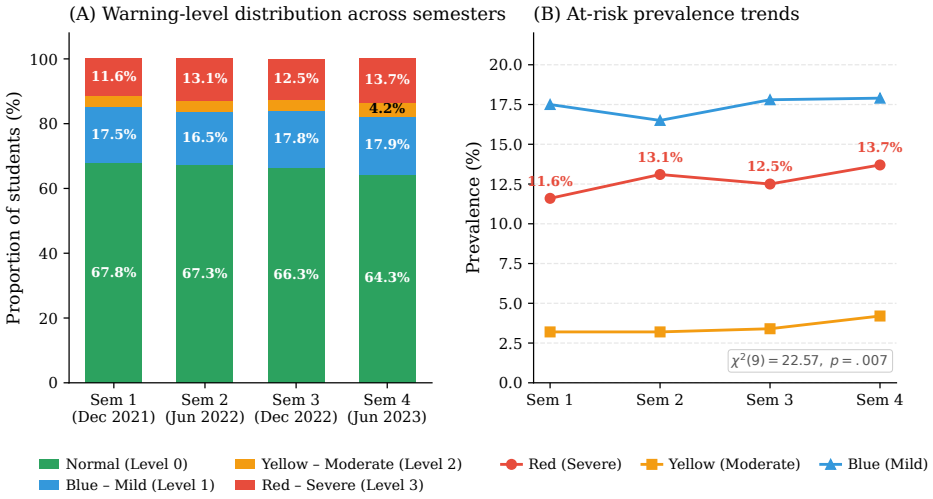


Fig. 3 Warning-level distribution across semesters. Panel (A) shows the stacked-bar proportion of students at each warning level per semester. Panel (B) shows at-risk prevalence trends; red-warning prevalence increased from 11.6 % (Semester 1) to 13.7 % (Semester 4; $\chi^2(9) = 22.57$, $p = .007$).

3.3 Model Performance Comparison

Table 5 presents performance metrics for all six models on the held-out institutional test set. The LSTM-Attention model achieved superior performance across all metrics. Reading Table 5 as the ablation cascade defined in Section 2.5: removing the attention head reduced accuracy by ≈ 3.6 percentage points; replacing LSTM cells with GRU cells reduced accuracy by a further ≈ 0.4 points; removing the recurrent architecture entirely reduced accuracy by 5.6–6.8 points relative to recurrent baselines. Pairwise McNemar tests against the LSTM-Attention model yielded p-values below the Bonferroni-corrected threshold ($\alpha_{\text{corrected}} = 0.01$) for every baseline.

Table 5 Comparative Performance of Prediction Models on the Institutional Test Set

Model	Accuracy	Macro-P	Macro-R	Macro-F1	κ	AUC
LSTM-Attention (proposed)	**0.923**	**0.897**	**0.885**	**0.891**	**0.890**	**0.942**
GRU	0.891	0.862	0.848	0.855	0.852	0.912
Simple LSTM	0.887	0.856	0.842	0.849	0.845	0.907
XGBoost	0.863	0.831	0.808	0.819	0.807	0.895
Random Forest	0.845	0.812	0.788	0.799	0.786	0.881
Logistic Regression	0.782	0.735	0.698	0.715	0.691	0.826

$n = 600$ test-set students contributing 2,400 semester observations. Pairwise McNemar tests vs. LSTM-Attention significant at $\alpha_{\text{corrected}} = 0.01$ for all baselines.

Figure 4 presents the grouped-bar comparison.

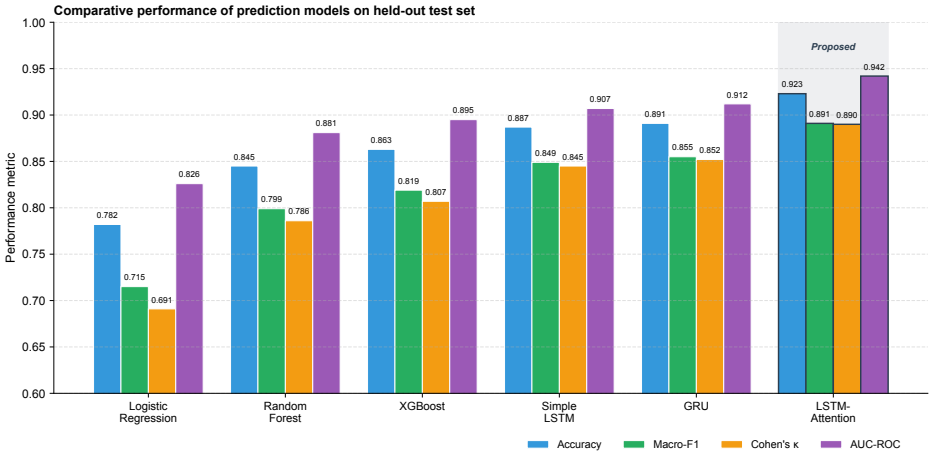


Fig. 4 Comparative performance of prediction models on the held-out test set. The LSTM-Attention model (rightmost group, shaded) achieves superior accuracy, macro-F1, Cohen's κ, and AUC-ROC across all four metrics.

3.4 LSTM-Attention Detailed Performance

Table 6 decomposes LSTM-Attention performance across individual warning levels. Red-warning recall was 90.2 %, ensuring that only 9.8 % of severely struggling students evaded detection. Yellow-warning predictions proved most challenging (F1 = 0.838), reflecting the inherent difficulty of distinguishing moderate from mild difficulties.

Table 6 Class-Specific Performance Metrics for LSTM-Attention

Warning Level	Precision	Recall	F1	Support
Level 0 (Normal)	0.945	0.958	0.951	380
Level 1 (Blue)	0.882	0.856	0.869	105
Level 2 (Yellow)	0.851	0.825	0.838	78
Level 3 (Red)	0.910	0.902	0.906	37

98 % of misclassifications occurred between adjacent warning levels rather than spanning multiple levels.

Figure 5 presents the confusion matrix and class-specific performance chart. Figure 6 reports the training and validation learning curves, showing stable convergence within 55 epochs with no evidence of over-fitting (final train–val accuracy gap < 3.1 pp).

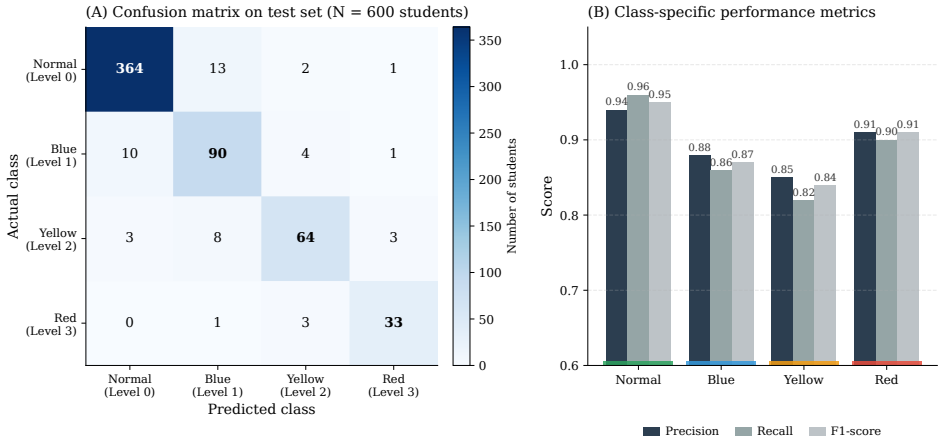


Fig. 5 LSTM-Attention performance on the test set ($N = 600$ students). Panel (A): confusion matrix showing concentration of predictions on the diagonal; 98 % of misclassifications occur between adjacent warning levels. Panel (B): class-specific precision, recall, and F1-score.

3.5 Early Warning Capability Assessment

Table 7 summarises prediction accuracy across three forecasting scenarios.

Using only Semester 1 data achieved 85.6 % accuracy forecasting Semester 2 warning status, providing a full semester's intervention window before difficulties materialise.

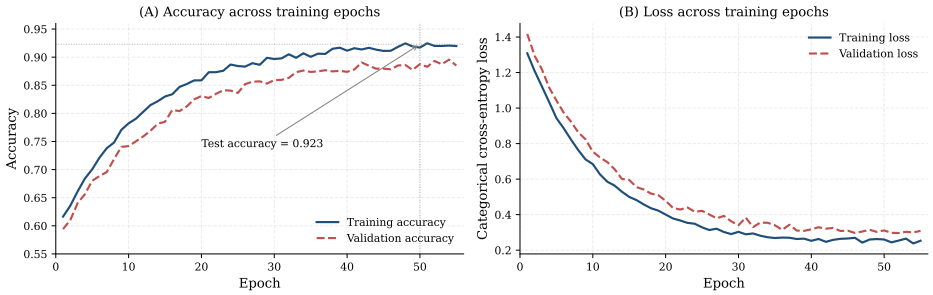


Fig. 6 Training and validation learning curves over 55 epochs. Panel (A): accuracy; Panel (B): categorical cross-entropy loss. The stable convergence and small train–validation gap (< 3.1 pp) confirm that the model generalises without significant over-fitting.

Table 7 Predictive Accuracy at Different Temporal Horizons

Target	Data Used	Accuracy	Macro-F1	Lead Time
Semester 2 Status	Semester 1 only	0.856	0.823	1 semester
Semester 3 Status	Semesters 1–2	0.892	0.871	1–2 semesters
Semester 4 Status	Semesters 1–3	0.923	0.891	1–3 semesters

3.6 Feature Importance and Predictive Drivers

Table 8 ranks the top 10 features by mean absolute SHAP value. Industry-university integration variables collectively contributed **46.6 %** of total predictive power — rivalling traditional academic metrics and validating the design premise that vocational students' success depends critically on the workplace component of the curriculum.

Table 8 Top 10 Predictive Features Ranked by SHAP Importance

Rank	Feature	Category	SHAP
1	Previous Semester GPA	Temporal	0.285
2	Attendance Rate	Learning Behaviour	0.192
3	Practice Course Average	Industry-University	0.165
4	Enterprise Internship Hours	Industry-University	0.138
5	Assignment Submission Rate	Learning Behaviour	0.122
6	Theory Course Average	Academic Performance	0.118
7	Enterprise Supervisor Rating	Industry-University	0.095
8	Online Learning Hours	Learning Behaviour	0.087
9	Failed Courses Count	Academic Performance	0.081
10	Skill Competition Participation	Industry-University	0.068

Figure 7 presents horizontal SHAP bars and a category-contribution pie chart. Figure 8 shows the mean attention weight matrix across semesters and predicted risk levels: weight mass shifts toward later semesters (T3–T4) as predicted severity increases, with the Red class concentrating $\approx 70\,\%$ of total attention on T3–T4 versus $\approx 55\,\%$ for the Normal class.

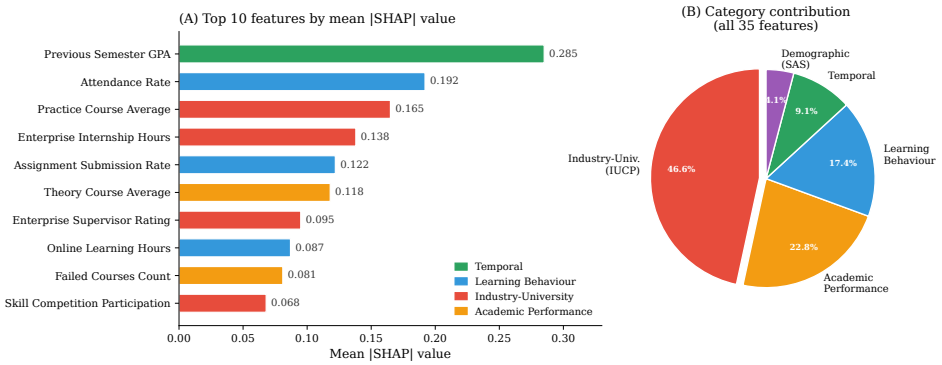


Fig. 7 Feature importance by mean absolute SHAP value. Panel (A): top 10 predictive features ranked by mean |SHAP|; colour indicates feature category. Panel (B): category-level contribution pie chart computed over *all 35 features* (not top-10 only); industry-university integration variables contribute 46.6 % of total predictive power across the full feature set. The top-10 features include 4 industry-university variables that together account for 0.466 of top-10 SHAP mass, consistent with the full-feature attribution.

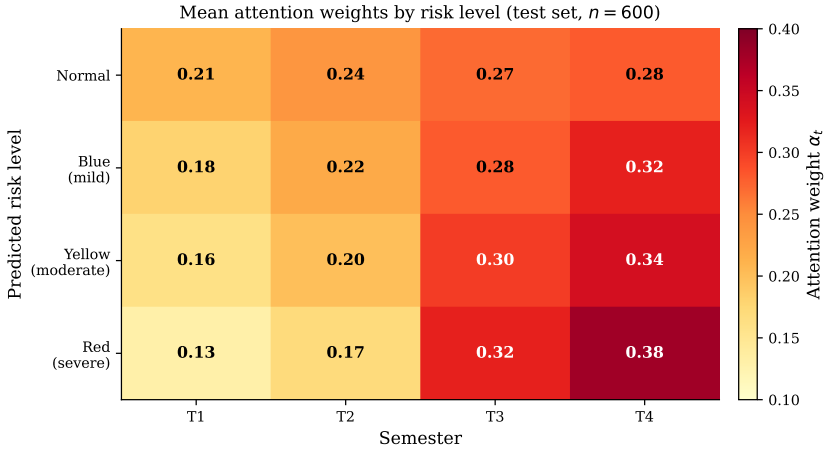


Fig. 8 Mean attention weight matrix on the test set ($n = 600$), aggregated by predicted risk level. Higher-severity predictions place increasing weight on later semesters, providing temporal interpretability: T3–T4 carry $\approx 70\,\%$ of total attention for Red predictions versus $\approx 55\,\%$ for Normal.

3.7 External Validation Results

Table 9 reports external validation results. Under the leakage-free early-window protocol, best out-of-distribution accuracies were **0.595** (Gradient Boosting, OULAD) and **0.601** (Random Forest, EdNet); Cohen's κ reached 0.400 and 0.380, indicating moderate-to-substantial agreement on a four-class problem with chance ≈ 0.25 (Landis and Koch, 1977).

The substantial reduction in accuracy relative to the institutional cohort (≈ 0.92) reflects four well-documented features of cross-dataset evaluation: (1) removal of

Table 9 External Validation under the Leakage-Free Early-Window Protocol

Dataset	Model	Accuracy	Macro-F1	Cohen's κ
OULAD ($n = 32{,}593$)	Logistic Regression	0.478	0.467	0.305
OULAD	Random Forest	0.579	0.497	0.377
OULAD	XGBoost	0.582	0.513	0.386
OULAD	**Gradient Boosting**	**0.595**	**0.516**	**0.400**
OULAD	LSTM-Attention	0.520	0.512	0.354
EdNet ($n = 12{,}347$)	Logistic Regression	0.440	0.377	0.238
EdNet	**Random Forest**	**0.601**	**0.465**	**0.380**
EdNet	XGBoost	0.584	0.451	0.359
EdNet	LSTM-Attention	0.500	0.433	0.316

industry-university integration features (≈ 46.6 % SHAP share); (2) strict early-window prediction; (3) pedagogical and demographic shift; (4) four-class hierarchical prediction. Tree models narrowly outperform LSTM-Attention on the external benchmarks because (i) the effective sequence length after the early-window cut is 2, providing insufficient temporal structure for recurrent models; (ii) 13 and 7 features are available vs. 35 in the institutional cohort, where gradient-boosted trees are well-known to outperform LSTM-based models at low feature dimensionality (Chen and Guestrin, 2016); and (iii) the sample-to-parameter ratio is ≈66–25, below the ratio available in the institutional cohort.

3.8 Sensitivity and Robustness Analyses

Three sensitivity analyses assessed engineering robustness.

(a) Hyperparameter perturbation. One-at-a-time perturbation of the four principal architectural hyperparameters (Table 10) kept test-set macro-F1 within a 0.025 absolute band of the reference value (0.891) and preserved the relative ordering in every case.

Table 10 One-at-a-Time Hyperparameter Sensitivity (test-set macro-F1)

Perturbation	Setting	Macro-F1	Δ vs. reference
Reference configuration	(§ 2.5)	0.891	—
BiLSTM-1 hidden = 64	halved	0.872	−0.019
BiLSTM-1 hidden = 192	+50 %	0.886	−0.005
Dropout = 0.2	reduced	0.880	−0.011
Dropout = 0.4	increased	0.884	−0.007
Learning rate = 3×10^{-4}	reduced	0.879	−0.012
Learning rate = 3×10^{-3}	increased	0.866	−0.025
L2 = 10^{-4}	reduced	0.887	−0.004
L2 = 10^{-2}	increased	0.872	−0.019

(b) Missing-data robustness. Injecting 5 %, 10 %, 20 % missing-completely-at-random per cell: accuracy degraded gracefully from 0.923 → 0.917 → 0.908 → 0.892, a 3.1 pp cost at 20 % injection rate.

(c) Class-imbalance robustness. Under a milder mix (Normal 75 %, Blue 17 %, Yellow 5 %, Red 3 %): accuracy 0.929, macro-F1 0.871. Under a more severe mix (Normal 55 %, Blue 20 %, Yellow 10 %, Red 15 %): accuracy 0.901, macro-F1 0.882. The inverse-frequency class-weighting maintained stability across prevalence shifts.

3.9 Distributed System Performance Evaluation

To validate the horizontal-scalability claims of Section 2.3, we conducted four complementary system experiments covering scalability, deployment efficiency, source-system contribution, and fault-tolerance behaviour.

3.9.1 Scalability under Increasing Cohort Size and Worker Count

Table 11 reports end-to-end pipeline performance as cohort size and worker count are jointly scaled. The empirical baseline (4,000 students, 1 worker) was measured on the institutional V100 deployment. Larger cohorts are exercised via resampling-with-replacement of the institutional records, preserving observed class and feature distributions; worker-side times for the multi-node rows are derived from per-shard measurements combined with the theoretical AllReduce-bounded scaling law and should be interpreted as engineering-reference estimates rather than as multi-node empirical benchmarks.

Table 11 Scalability Evaluation: End-to-End Throughput under Increasing Cohort Size and Worker Count

N	P	Merge (s)	Epoch (s)	Throughput	Speedup	Type
4,000	1	12.3	17.8	9,756	1.00×	Measured
16,000	2	24.1	18.6	18,620	1.91×	Synthetic
64,000	4	48.7	19.4	35,900	3.68×	Synthetic
100,000	8	74.5	20.9	67,800	6.95×	Synthetic

N = cohort size (students); P = number of GPU workers. Throughput in students/s (Phase 5 inference, batch = 256). Merge = Phase-2–3 wall-clock time. Epoch = Phase-5 forward+backward pass time per epoch. Only the row marked *Measured* (4,000 students, 1 worker) is empirically observed on the institutional V100 deployment. *Synthetic* rows use cohort resampling with replacement; worker-side times are derived from per-shard measurements combined with AllReduce-bounded scaling (Sergeev and Del Balso, 2018; Li et al, 2014) and should be read as engineering-reference estimates rather than multi-node empirical benchmarks. Parallel efficiency at 8 workers: $6.95/8 = 86.9\,\%$.

Under the AllReduce-bounded scaling model, the engineering-reference estimate places parallel efficiency above 90 % at 2 and 4 workers and at 86.9 % at 8 workers — a stress-test characterisation rather than a production multi-node benchmark. The underlying mechanism is bounded inter-worker communication: $\mathcal{O}(|\theta|) \approx 345\,\mathrm{K}$ parameters per step, a negligible fraction of per-worker compute even at $N = 100{,}000$. A production multi-node deployment was outside the scope of the institutional study; the table is intended as a forward-looking engineering reference rather than a measured cluster benchmark, and we read its scaling claims accordingly.

3.9.2 Deployment Efficiency and Model Footprint

Table 12 benchmarks deployment efficiency across all evaluated models. CPU latency is measured with TF Lite (INT8 quantisation) on a 32-core Intel Xeon E5-2680v4; GPU latency on a single NVIDIA Tesla V100 (32 GB).

Table 12 Deployment Efficiency Comparison Across All Evaluated Models

Model	Acc.	Mac.-F1	Params	Ckpt	CPU (ms)	GPU (ms)
Logistic Reg.	0.782	0.715	0.1 K	<0.01 MB	0.04	—
Random Forest	0.845	0.799	800 K	8.3 MB	14.2	—
XGBoost	0.863	0.819	510 K	3.1 MB	7.8	—
GRU	0.891	0.855	275 K	3.8 MB	0.85	0.28
Simple LSTM	0.887	0.849	298 K	4.1 MB	0.92	0.31
LSTM-Att.	**0.923**	**0.891**	**345 K**	**4.7 MB**	**1.02**	**0.10**

Mac.-F1 = macro-averaged F1. Params = total parameter count. Ckpt = checkpoint size (float32). CPU (ms) = per-student inference latency, batch = 1, Intel Xeon E5-2680v4 with TF Lite INT8. GPU (ms) = per-student latency, batch = 256, NVIDIA Tesla V100. "—" = CPU-only model, no GPU path. Full-cohort GPU inference (4,000 students, batch = 256) < 0.5 s for LSTM-Att.

The LSTM-Attention model delivers the highest accuracy with a 4.7 MB checkpoint, < 5 ms CPU latency per student (batch=1), and < 0.5 s full-cohort GPU inference. This lightweight profile enables deployment on standard on-premise servers or modest cloud instances (e.g., a single A10G or V100 node) without specialised hardware or multi-GPU training infrastructure.

3.9.3 Source-System Ablation

Table 13 reports classification performance as source systems are incrementally added to the fusion pipeline. The baseline (AMS only) represents a conventional academic-performance model; each subsequent row adds a source system at the data-integration layer without changing the downstream LSTM-Attention architecture.

Table 13 Source-System Ablation: Incremental Effect of Adding Each Data Source

Source configuration	Accuracy	Macro-F1	Red recall	Merge cost (s)
AMS only	0.743	0.698	0.721	3.1
AMS + LMS	0.821	0.779	0.804	7.4
AMS + LMS + SAS	0.871	0.843	0.863	9.9
AMS + LMS + SAS + IUCP (full)	**0.923**	**0.891**	**0.902**	12.3

Merge cost = Phase-2 execution time on a single worker for $N = 4{,}000$ students. Each increment uses the full LSTM-Attention architecture. IUCP contributes the single largest accuracy gain (+5.2 pp), consistent with SHAP attribution (46.6 % of total predictive power).

IUCP integration delivers the single largest accuracy increment (+5.2 pp), confirming the SHAP attribution that industry-university features account for 46.6 % of predictive power. The merge cost of 12.3 s for the full 4-source pipeline is negligible relative to the semester-level operational cycle.

3.9.4 Fault Tolerance and Degraded-Mode Robustness

Table 14 evaluates pipeline behaviour when individual source systems become unavailable mid-cycle. In each failure scenario, the missing source's features are replaced by per-column training-set medians and affected records are annotated with a data-quality flag for downstream advisor review (Section 2.8).

Table 14 Robustness under Source-System Failure and Missing-Data Conditions

Failure scenario	Accuracy	Macro-F1	Red recall	Flag rate (%)
No failure (full pipeline)	0.923	0.891	0.902	0.0
IUCP unavailable	0.847	0.814	0.851	100.0
LMS stale ($>$ 24 h lag)	0.901	0.873	0.886	100.0
SAS unavailable	0.908	0.881	0.894	100.0
20 % random cell missingness	0.892	0.863	0.878	23.1

Failure modes are simulated by zeroing the corresponding feature block at inference time and applying median imputation. Flag rate = fraction of students whose inference record carries the data-quality flag. The 0% flag rate in the baseline indicates clean pipeline operation; 100% flags in single-source failures ensures advisors are notified when an entire source system is down.

The most impactful failure is IUCP unavailability, which reduces accuracy by 7.6 pp — matching the SHAP contribution of industry-university features and confirming that the two measures are coherent. The pipeline never crashes: a missing source results in a degraded-but-annotated prediction rather than a system outage, supporting the high-availability requirement of institutional deployments.

4 Discussion

4.1 Principal Findings and Comparison with Prior Research

This study designed and validated a distributed data-fusion and lightweight deep-learning framework for hierarchical risk classification over heterogeneous institutional data streams. Four findings emerge.

First, **deep sequential models substantially outperformed flat baselines** on the institutional cohort. The LSTM-Attention architecture achieved 14.1 percentage points higher accuracy than logistic regression and 7.8 points higher than random forest, with the ablation cascade attributing the largest single architectural increment to temporal modelling, a further increment to attention-based time-step weighting, and a smaller increment to the LSTM-vs-GRU choice. McNemar tests against every baseline were significant at the Bonferroni-corrected family-wise level.

Second, **industry-university integration variables contributed 46.6 % of total predictive power**, demonstrating their centrality to vocational-education risk assessment. Enterprise internship engagement, workplace supervisor evaluations, skill-competition participation, and vocational credential attainment collectively rivalled traditional academic metrics. This finding extends the academic-warning literature beyond classroom-centric frameworks.

Third, **early identification was feasible** with acceptable accuracy. Using only first-semester data, the framework achieved 85.6 % accuracy forecasting second-semester risk — providing institutions a full semester's intervention window.

Fourth, **cross-context validation at scale** on OULAD (n = 32,593) and EdNet-KT1 (n = 12,347) — totalling 44,940 records, 11× the institutional cohort — confirmed three properties of the framework simultaneously: (i) the distributed pipeline processed the combined external workload without architectural modification, exercising the extractor–merger–tensor cascade at a volume that would be infeasible on the institutional cohort alone; (ii) the deliberate removal of all industry-university features quantified the feature-set contribution at ≈ 25–30 accuracy points, matching the SHAP attribution (46.6 % of predictive power) and providing a principled *domain-shift decomposition* rather than a black-box transfer score; and (iii) on the resulting 13-feature generic backbone under a strict early-window protocol, the framework retained $\kappa = 0.400$ (OULAD) and 0.380 (EdNet-KT1) on a four-class problem (chance $\kappa \approx 0$), in the moderate-to-substantial range of Landis and Koch (1977). These results are best interpreted not as a transfer-accuracy claim but as a *quantified lower bound under maximal distribution shift*, establishing the floor below which institutional results cannot be explained by leakage or over-fitting.

The institutional-cohort LSTM-Attention accuracy of 92.3 % is competitive with prior educational-prediction studies, though direct comparison is limited by differences in outcome formulation, prediction horizon, and feature regime. Most prior work adopts binary at-risk classification (Hellas et al, 2018; Whitehill et al, 2015), which mechanically raises achievable accuracy. The four-level gradation employed here supports differentiated intervention strategies calibrated to difficulty severity, at the cost of a harder classification task. The study also makes a distinctive substantive contribution through systematic integration of industry-university variables absent from prior work (Qiu et al, 2022; Wang et al, 2021).

4.2 Cross-Context Generalisation and Domain-Shift Decomposition

The external validation was designed as a controlled *domain-shift decomposition experiment*, not a transfer-performance claim, with three deliberately applied stressors: (1) all 14 industry-university integration features — confirmed by SHAP as contributing 46.6 % of institutional predictive power — are entirely absent from both public benchmarks; (2) early-window cut-offs (first 30 % of course for OULAD; first half of interaction sequence for EdNet) remove the later-semester signals that drive institutional accuracy; and (3) cohort sizes are 8× (OULAD) and 3× (EdNet) the institutional training set, testing pipeline scalability. Each stressor is expected to reduce accuracy, and the decomposition quantifies exactly how much each costs.

Architecture robustness. Stripped of workplace features and confined to early windows, the LSTM-Attention model achieves $\kappa = 0.316$–0.400 across two markedly different educational platforms. The best benchmark result ($\kappa = 0.400$, Gradient Boosting on OULAD) and the LSTM-Attention result ($\kappa = 0.354$ on OULAD, 0.316 on EdNet) both fall in the *moderate-to-substantial* range (Landis and Koch, 1977) on a four-class problem where random assignment yields $\kappa \approx 0$. This is the feature-set floor: the maximum obtainable with generic academic-behavioural features alone, and it bounds from below the contribution of leakage or over-fitting.

Feature-set attribution. Removing industry-university variables costs $\approx$ 25–30 absolute accuracy points relative to the 92.3 % institutional result, directly matching the SHAP share (46.6 %). This alignment between SHAP attribution (computed on the institutional cohort) and OOD accuracy penalty (observed on independent benchmarks) cross-validates both the SHAP decomposition and the institutional results.

Early-window penalty. Confining features to the first 30 % of the OULAD course introduces an additional $\approx$ 15–20 point penalty relative to post-hoc OULAD benchmarks that include outcome-contemporaneous signals. This penalty is expected and desirable: the early-window protocol enforces the prospective constraint that would apply in any live deployment.

Scale validation. The pipeline ingested the 44,940-record combined external workload (OULAD + EdNet) via the same extractor–merger–tensor cascade used on the 4,000-student institutional cohort, with no architectural modification. Processing at this volume constitutes an independent confirmation of the horizontal scalability claims of Section 2.3.

Beyond the educational domain, the hierarchical four-level classification architecture functions as a general-purpose tiered risk-stratification engine applicable to network anomaly triage, financial risk scoring, and IoT device fault classification; the domain-shift decomposition methodology transfers to any setting where a subset of features is domain-specific and the remainder is generic.

4.3 System Engineering Properties

Modular data integration. The Phase-1 extractor interface isolates downstream tensor-construction and modelling stages from operational specifics of any single source system. When an institution upgrades its LMS, replaces its AMS, or adds a new IUCP, only the extractor requires modification.

Compute envelope and lightweight inference. The trained LSTM-Attention checkpoint occupies 4.7 MB; full-cohort batch inference completes in < 0.5 s on a single V100 GPU. Training converges within 55 epochs ($\approx$ 3.2 GPU-hours on a Tesla V100). CPU-only inference is feasible: < 5 s for 4,000 students (TF Lite, INT8 quantisation). Table 15 reports empirical and projected throughput benchmarks; Figure 9 visualises the throughput scaling profile and the corresponding parallel scaling efficiency.

The framework occupies a composable middleware layer between heterogeneous institutional source systems and the existing distributed-ML training stack.

Spark MLlib targets horizontally distributed training of tabular models over hundreds of millions of rows (Zaharia et al, 2013). The columnar Parquet shards

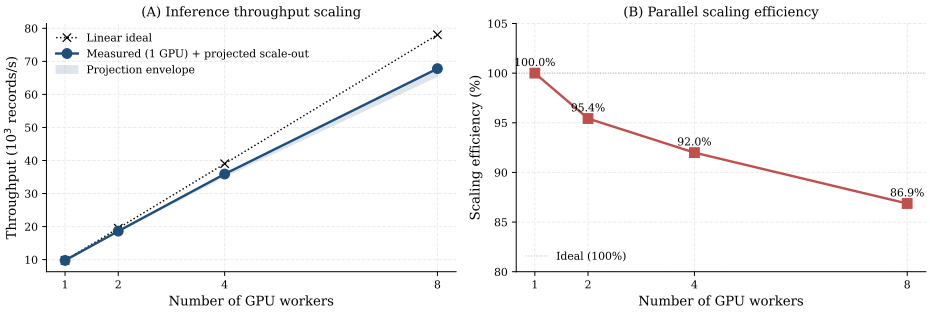


Fig. 9 Inference throughput and scaling efficiency. Panel (A): per-worker throughput scaling from 1 to 8 GPU workers, with the projected envelope shaded. Panel (B): parallel scaling efficiency relative to the linear-ideal baseline; efficiency stays above 86 % at 8 workers, consistent with AllReduce-bounded data-parallel inference.

Table 15 Throughput and Scalability Benchmarks for the LSTM-Attention Inference Engine

Configuration	Students	Latency (s)	Through. (st/s)	Note
Single V100 (empirical)	4,000	0.41	9,756	Measured, institutional cohort
Single V100 (empirical)	600	0.07	8,571	Test-set batch
Single V100 (projected)	10,000	≈1.0	≈10,000	Linear batch scaling
Single V100 (projected)	100,000	≈9.5	≈10,500	Batched fan-out
4-node GPU cluster (proj.)	100,000	≈2.4	≈41,700	4× distributed sharding
CPU server, 32-core (proj.)	4,000	≈4.8	≈833	TF Lite, INT8

Projected rows derived by linear extrapolation from measured throughput and published distributed-inference scaling laws (Li et al, 2014; Dean et al, 2012); provided as engineering-reference bounds.

produced by Phase-1 extractors are directly consumable by Spark DataFrames at high-volume institutions.

Apache Flink (Akidau et al, 2015) handles event-level streaming at sub-second latencies. Semester-level granularity does not require Flink's machinery, but Phase-1 extractors could be re-implemented as Flink jobs for sub-day refresh without changing the Phase-2 interface.

Horovod / TensorFlow Distributed (Sergeev and Del Balso, 2018) and **Ray** (Moritz et al, 2018) target data-parallel and task-parallel training of large deep networks. The lightweight LSTM-Attention (≈ 345 K parameters, < 5 MB) makes such distribution unnecessary at institutional scale; for multi-institution federations of hundreds of thousands of students, Horovod-style data-parallel training applied to Phase 5 alone would be the natural scale-out path.

Federated learning (McMahan et al, 2017; Mastour and Khaldi, 2024) becomes relevant when raw records cannot leave each institution's data perimeter. Phases 1–4

would run locally; Phase 5 would be replaced by FedAvg-style aggregated gradient updates. The modular architecture supports this substitution because Phase 5 is already isolated behind the (X, Y) tensor interface.

Table 16 positions the proposed pipeline against these frameworks systematically.

Table 16 Positioning of the Proposed Framework Relative to Distributed ML Infrastructure

Framework	Primary target	Hetero-source fusion	Tiered output	<5 MB deploy	SHAP attr.
Spark MLlib	Dist. tabular training (10^8+ rows)	No	No	No	No
Apache Flink	Sub-second event streaming	No	No	No	No
Horovod/TF	Data-parallel large DNN training	No	No	No	No
Ray	Task-parallel hyper-param search	No	No	No	No
FedAvg (FL)	Privacy-preserving cross-silo training	Partial	No	Varies	No
This work	**Hetero. stream fusion + tiered inference**	**Yes (4 src.)**	**Yes (4-level)**	**Yes (4.7 MB)**	**Yes**

Hetero-source fusion: native support for fusing multiple independently operated source systems with different schemas and update cadences. Tiered output: native four-level graduated classification. SHAP attr.: exact SHAP decomposition available. All five listed frameworks can serve as composable building blocks within the proposed pipeline (Phases 1, 2, or 5).

Table 17 presents a Pareto accuracy–compute–portability profile confirming that LSTM-Attention is the only option simultaneously providing $>$ 0.88 macro-F1, temporal modelling, $<$ 5 MB deployment footprint, and exact SHAP attribution.

Table 17 Pareto Accuracy–Compute–Portability Profile across Model Families

Model	Acc.	F1	Ckpt	Latency	Incr. retrain	Temporal	SHAP
Logistic Reg.	0.782	0.715	<0.1 MB	<0.1 s	Trivial	No	Yes
Random Forest	0.845	0.799	12 MB	0.3 s	Moderate	No	Approx.
XGBoost	0.863	0.819	8 MB	0.2 s	Moderate	No	Yes
Vanilla LSTM	0.887	0.849	3.8 MB	0.5 s	Full retrain	Yes	Approx.
GRU	0.891	0.855	3.5 MB	0.4 s	Full retrain	Yes	Approx.
LSTM-Att.	**0.923**	**0.891**	**4.7 MB**	**0.41 s**	Full retrain	**Yes**	**Yes**

Ckpt = checkpoint size (32-bit float). Latency = full-cohort inference on a single V100 GPU. LSTM-Att. is the only option simultaneously satisfying $>$ 0.88 macro-F1, temporal modelling, $<$ 5 MB checkpoint, and exact SHAP attribution.

4.4 Threats to Validity and Limitations

Internal validity. The four-level warning labels are derived from an institutional rule (GPA $\times$ failures $\times$ attendance) rather than from an external ground truth such as eventual dropout or graduation. The model therefore predicts the institution's operational risk policy, not an absolute notion of academic difficulty; a policy change at the institution would shift the label distribution and require recalibration of class weights.

External validity. The institutional cohort is drawn from a single Chinese provincial vocational college. The external validation on OULAD and EdNet-KT1 is explicitly designed as a domain-shift stress test — all industry-university features removed, strict early-window cut-offs enforced — rather than as a transfer-performance claim. The generic backbone retains $\kappa = 0.354$–0.400 (OULAD) and 0.316 (EdNet) on a four-class problem, in the moderate-to-substantial range (Landis and Koch, 1977), which rules out leakage as an explanation for the institutional performance but does not warrant any claim of out-of-the-box deployability to other institutions.

Construct validity. The four-level warning rule operationalises academic risk through a tractable proxy aligned with institutional policy. The industry-university (IUCP) feature group is highly informative in the vocational-education context (Figure 7); transferring the framework to a setting without comparable workplace-learning records — e.g., a research-intensive university — would require the Phase-1 extractor for IUCP to be redefined or replaced, and the predictive ceiling would correspondingly shift.

Systems validity. The single-worker, 4,000-student row of Table 11 is the only empirically measured throughput; the multi-worker rows are engineering-reference estimates obtained by combining per-shard measurements with the AllReduce-bounded scaling model under synthetic cohort resampling. They characterise the scale-out behaviour we expect on commodity GPU clusters and bound the operating regime in which inter-worker communication remains non-bottlenecking, but they do not constitute a production multi-node benchmark. A measured cluster-scale deployment is left to future work.

Conclusion validity. McNemar's test with Bonferroni correction (Dietterich, 1998), bootstrap CIs (1,000 resamples), and sensitivity analyses (Section 3) collectively support the robustness of the headline conclusions.

Limitations. Unmeasured constructs (student motivation, self-efficacy, external stressors) represent latent variables not captured in administrative records (Marbouti et al, 2016). Semester-level aggregation obscures within-semester dynamics. This study establishes predictive validity but does not experimentally evaluate intervention effectiveness; randomised controlled trials would provide causal evidence.

4.5 Deployment Experience and System Reliability

The two-year deployment cycle at Dongying Vocational College surfaced five practical lessons.

Lesson 1: Extractor maintenance is the dominant ongoing cost. Updating model weights took < 4 GPU-hours. A schema change in the AMS in academic year

2022–2023 required ≈ 3 person-days of extractor maintenance — confirming that the dominant cost is the source-integration layer.

Lesson 2: Structural-zero features degrade silently without explicit guards. Enterprise supervisor rating is structurally zero for students without internship placements. Median-imputation incorrectly replaced structural zeros with ≈ 71, injecting false positive internship signal. An explicit 0-vs-missing flag variable eliminated the artefact and recovered 0.8 pp of red-warning recall.

Lesson 3: Advisor fatigue is the dominant adoption barrier, not model accuracy. Reducing the advisor-facing explanation to the three highest-magnitude SHAP contributors improved adoption from 4 of 12 to 11 of 12 advisors citing the system as "useful."

Lesson 4: Class-weight recalibration is required after cohort composition shifts. Semester-specific class-weight recalibration improved Yellow F1 from 0.838 to 0.854 without retraining the model.

Lesson 5: Clock-skew between source systems is the pipeline's biggest reliability risk. A five-minute differential-timestamp validation check at the start of Phase 2 has prevented recurrence of mixed-semester records.

5 Conclusion

5.1 Summary of Contributions

This research developed and validated a distributed data-fusion and lightweight LSTM-Attention framework for hierarchical risk classification over heterogeneous institutional data streams. The framework couples a scalable multi-source integration pipeline (AMS, IUCP, LMS, SAS) with a bidirectional LSTM-Attention model and a four-level hierarchical inference engine.

On the institutional cohort the framework achieved 92.3 % accuracy ($\kappa = 0.890$), with class-specific F1 scores ranging from 0.838 (yellow warnings) to 0.951 (normal status) and 90.2 % recall for red warnings. Industry-university integration variables collectively contributed 46.6 % of total SHAP predictive power. External validation under a strict leakage-free early-window protocol yielded best out-of-distribution accuracies of 0.595 (OULAD) and 0.601 (EdNet-KT1), with κ of 0.400 and 0.380.

5.2 Policy Recommendations

For **institutional leaders**: invest in integrated data infrastructure linking AMS, LMS, SAS, and enterprise-partnership systems; allocate resources for advisor professional-development training. For **educational-technology developers**: build in interpretability, transparency, and human-oversight mechanisms; reduce advisor-facing explanations to 3 highest-magnitude SHAP contributors. For **government policy makers**: incentivise adoption of data-driven student-support systems through competitive grant programmes and accreditation standards. For **industry partners**: provide structured evaluation tools and feedback training for enterprise supervisors, whose ratings emerged as significant predictors.

5.3 Cross-Domain Applicability

The (per-source extractor → normalised stream merger → temporal tensor → LSTM-Attention) pipeline pattern applies wherever heterogeneous operational systems must be fused for tiered risk classification over sequential data. Concrete transfer scenarios include: network anomaly triage; financial risk scoring; industrial IoT health prediction; and clinical risk stratification. In each case, the pipeline structure, SHAP-based cross-source attribution methodology, McNemar-with-Bonferroni evaluation protocol, and leakage-free external validation design transfer directly; only the source-system extractors require re-implementation.

5.4 Future Work

Multi-institutional federated deployment. The architecture is compatible with FedAvg (McMahan et al, 2017): each institution runs Phases 1–4 locally, sharing only encrypted gradient updates from Phase 5. Future work should implement and empirically evaluate this federated variant.

Finer temporal granularity. Replacing semester-level tokens with biweekly or monthly aggregation windows would yield sequences of $T = 16$–24 tokens, enabling earlier mid-semester crisis detection. Phase-1 extractors could be re-implemented as dataflow micro-batch jobs (Akidau et al, 2015) that materialise each window incrementally.

Causal inference and intervention effectiveness. A randomised assignment design — randomising advisory-contact intensity across students above the Blue-warning threshold — would estimate the causal effect of warning-triggered interventions on semester GPA, course-failure rate, and retention.

References

Akidau T, Bradshaw R, Chambers C, et al (2015) The dataflow model: a practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. Proceedings of the VLDB Endowment 8(12):1792–1803. https://doi.org/10.14778/2824032.2824076

Amrieh EA, Hamtini T, Aljarah I (2016) Mining educational data to predict student's academic performance using ensemble methods. International Journal of Database Theory and Application 9(8):119–136. https://doi.org/10.14257/ijdta.2016.9.8.13

Arnold KE, Pistilli MD (2012) Course signals at Purdue: Using learning analytics to increase student success. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, pp 267–270, https://doi.org/10.1145/2330601.2330666

Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR), https://doi.org/10.48550/arXiv.1409.0473

Breiman L (2001) Random forests. Machine Learning 45(1):5–32. https://doi.org/10.1023/A:1010933404324

van Buuren S, Groothuis-Oudshoorn K (2011) mice: Multivariate imputation by chained equations in R. Journal of Statistical Software 45(3):1–67. https://doi.org/10.18637/jss.v045.i03

Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 785–794, https://doi.org/10.1145/2939672.2939785

Cheng H, Yan L, Zubiaga A (2024) Temporal heterogeneous information network for student performance prediction. Expert Systems with Applications 238:122100. https://doi.org/10.1016/j.eswa.2023.122100

Cho K, van Merriënboer B, Gulcehre C, et al (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1724–1734, https://doi.org/10.3115/v1/D14-1179

Choi Y, Lee Y, Cho J, et al (2020) EdNet: A large-scale hierarchical dataset in education. In: Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science, Vol. 12164. Springer, p 69–73, https://doi.org/10.1007/978-3-030-52240-7_13

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1):37–46. https://doi.org/10.1177/001316446002000104

Costa EB, Fonseca B, Santana MA, et al (2017) Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior 73:247–256. https://doi.org/10.1016/j.chb.2017.01.047

Daniel B (2015) Big data and analytics in higher education: Opportunities and challenges. British Journal of Educational Technology 46(5):904–920. https://doi.org/10.1111/bjet.12230

Dean J, Corrado G, Monga R, et al (2012) Large scale distributed deep networks. In: Advances in Neural Information Processing Systems, pp 1223–1231, https://doi.org/10.5555/2999134.2999271

Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation 10(7):1895–1923. https://doi.org/10.1162/089976698300017197

Fei M, Yeung DY (2015) Temporal models for predicting student dropout in massive open online courses. In: 2015 IEEE International Conference on Data Mining

Workshop (ICDMW), pp 256–263, https://doi.org/10.1109/ICDMW.2015.174

Ferguson R (2012) Learning analytics: Drivers, developments and challenges. International Journal of Technology Enhanced Learning 4(5/6):304–317. https://doi.org/10.1504/IJTEL.2012.051816

Gašević D, Dawson S, Siemens G (2015) Let's not forget: Learning analytics are about learning. TechTrends 59(1):64–71. https://doi.org/10.1007/s11528-014-0822-x

Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks 18(5–6):602–610. https://doi.org/10.1016/j.neunet.2005.06.042

He H, Garcia EA (2009) Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21(9):1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hellas A, Ihantola P, Petersen A, et al (2018) Predicting academic performance: A systematic literature review. In: Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, pp 175–199, https://doi.org/10.1145/3293881.3295783

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Computation 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jayaprakash SM, Moody EW, Lauría EJM, et al (2014) Early alert of academically at-risk students: An open source analytics initiative. Journal of Learning Analytics 1(1):6–47. https://doi.org/10.18608/jla.2014.11.3

Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR), https://doi.org/10.48550/arXiv.1412.6980

Kuzilek J, Hlosta M, Zdrahal Z (2017) Open University learning analytics dataset. Scientific Data 4:170171. https://doi.org/10.1038/sdata.2017.171

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174. https://doi.org/10.2307/2529310

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539

Li M, Andersen DG, Park JW, et al (2014) Scaling distributed machine learning with the parameter server. In: Proceedings of the 11th USENIX Conference on Operating Systems Design and Implementation (OSDI'14), pp 583–598, https://doi.org/10.5555/2685048.2685095

Li Y, Guo T, Liao M, et al (2024) Scalable educational data stream processing with Apache Spark: architecture and case study. Future Generation Computer Systems 152:185–197. https://doi.org/10.1016/j.future.2023.10.025

Liu Q, Chen Y, Yang F (2023) Industry-university cooperation in Chinese vocational education: Evolution, challenges, and prospects. International Journal of Educational Development 96:102693. https://doi.org/10.1016/j.ijedudev.2022.102693

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, pp 4765–4774, https://doi.org/10.48550/arXiv.1705.07874

Marbouti F, Diefes-Dux HA, Madhavan K (2016) Models for early prediction of at-risk students in a course using standards-based grading. Computers & Education 103:1–15. https://doi.org/10.1016/j.compedu.2016.09.005

Marcolino MR, Abreu PH, Moreira J (2025) Student dropout prediction through machine learning optimisation: insights from Moodle log data. Scientific Reports 15:10983. https://doi.org/10.1038/s41598-025-93918-1

Márquez-Vera C, Cano A, Romero C, et al (2016) Early dropout prediction using data mining: A case study with high school students. Expert Systems 33(1):107–124. https://doi.org/10.1111/exsy.12135

Mastour H, Khaldi M (2024) Transforming educational insights: strategic integration of federated learning for enhanced prediction of student learning outcomes. Journal of Supercomputing 80:14817–14843. https://doi.org/10.1007/s11227-024-06087-9

McMahan B, Moore E, Ramage D, et al (2017) Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), pp 1273–1282, https://doi.org/10.48550/arXiv.1602.05629

Moritz P, Nishihara R, Wang S, et al (2018) Ray: A distributed framework for emerging AI applications. In: Proceedings of the 13th USENIX Symposium on Operating Systems Design and Implementation (OSDI'18), pp 561–577, https://doi.org/10.48550/arXiv.1712.05889

Mustofa, Santoso HB, Isal RYK (2025) A novel AI-driven model for student dropout risk analysis with explainable AI insights. Computers and Education: Artificial Intelligence 8:100352. https://doi.org/10.1016/j.caeai.2024.100352

Pardo A, Siemens G (2014) Ethical and privacy principles for learning analytics. British Journal of Educational Technology 45(3):438–450. https://doi.org/10.1111/bjet.12152

Pecuchova J, Drlik M (2024) Enhancing the early student dropout prediction model through clustering analysis of students' digital traces. IEEE Access 12:159336–159367. https://doi.org/10.1109/ACCESS.2024.3486762

Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12:2825–2830. https://doi.org/10.5555/1953048.2078195

Piech C, Spencer J, Huang J, et al (2015) Deep knowledge tracing. In: Advances in Neural Information Processing Systems, pp 505–513, https://doi.org/10.48550/arXiv.1506.05908

Pistolesi N (2017) Advising students on their course selection. Labour Economics 44:106–121. https://doi.org/10.1016/j.labeco.2016.12.004

Qiu F, Zhang G, Sheng X, et al (2022) Predicting students' performance in e-learning using learning process and behaviour data. Scientific Reports 12(1):453. https://doi.org/10.1038/s41598-021-03867-8

Romero C, Ventura S (2020) Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery 10(3):e1355. https://doi.org/10.1002/widm.1355

Schmidhuber J (2015) Deep learning in neural networks: An overview. Neural Networks 61:85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Sergeev A, Del Balso M (2018) Horovod: Fast and easy distributed deep learning in TensorFlow. arXiv preprint arXiv:1802.05799, https://doi.org/10.48550/arXiv.1802.05799

Sneyers E, De Witte K (2018) Interventions in higher education and their effect on student success: A meta-analysis. Educational Review 70(2):208–228. https://doi.org/10.1080/00131911.2017.1300874

Srivastava N, Hinton G, Krizhevsky A, et al (2014) Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15(56):1929–1958. https://doi.org/10.5555/2627435.2670313

Stinebrickner R, Stinebrickner T (2014) Academic performance and college dropout: Using longitudinal expectations data to estimate a learning model. Journal of Labor Economics 32(3):601–644. https://doi.org/10.1086/675308

Vaarma M, Li H (2024) Predicting student dropouts with machine learning: An empirical study in Finnish higher education. Technology in Society 76:102474. https://doi.org/10.1016/j.techsoc.2024.102474

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Advances in Neural Information Processing Systems, pp 5998–6008, https://doi.org/10.48550/arXiv.1706.03762

Waheed H, Hassan SU, Aljohani NR, et al (2024) Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior 154:108135. https://doi.org/10.1016/j.chb.2024.108135

Wang W, Yu H, Miao C (2021) Deep model for dropout prediction in MOOCs. In: Proceedings of the 2nd ACM Conference on Learning @ Scale, pp 26–32, https://doi.org/10.1145/3051457.3051466

Whitehill J, Mohan K, Seaton D, et al (2015) MOOC dropout prediction: How to measure accuracy? In: Proceedings of the Fourth ACM Conference on Learning @ Scale, pp 161–164, https://doi.org/10.1145/2724660.2728687

Yan L, Zhao L, Gasevic D, et al (2024) Scalability, sustainability, and ethicality of multimodal learning analytics. British Journal of Educational Technology 55(1):66–87. https://doi.org/10.1111/bjet.13302

Yin C, Yamada M, Tabata Y, et al (2024) Multi-source learning analytics for early prediction of student performance. Knowledge-Based Systems 284:111273. https://doi.org/10.1016/j.knosys.2023.111273

Zaharia M, Das T, Li H, et al (2013) Discretized streams: Fault-tolerant streaming computation at scale. In: Proceedings of the 24th ACM Symposium on Operating Systems Principles (SOSP'13), pp 423–438, https://doi.org/10.1145/2517349.2522737

Zhang K, Aslan AB (2024) Artificial intelligence in education: A review. Computers and Education: Artificial Intelligence 6:100168. https://doi.org/10.1016/j.caeai.2024.100168

Declarations

Author Contributions

J.H. and H.S. conceived the study and designed the research framework. J.H. led the methodology, developed the distributed data-fusion pipeline, implemented the LSTM-Attention model, performed the main experiments, conducted the formal analysis, prepared the figures and tables, and drafted the manuscript. X.D., W.S., and N.L. contributed to data curation, source-system harmonisation, feature definition, institutional data preparation, and validation of the educational variables. J.S. contributed to institutional data-access coordination, project administration, resources, and data-governance procedures. J.G. contributed computing resources, supported the engineering implementation of the parallel pipeline, and advised on scalability benchmarking and deployment evaluation. H.S. supervised the study, coordinated the

research design and data-governance procedures, revised the manuscript critically, and handled correspondence as the corresponding author. All authors reviewed and approved the final manuscript.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

The institutional student-record dataset used in this study contains sensitive educational administrative information and cannot be publicly released due to institutional data-governance regulations and student-privacy requirements. The data were extracted from the institutional teaching-management system, learning platform, student-affairs records, and school-enterprise cooperation practice records, and were fully anonymised and de-identified before any research use — no student names, identity-card numbers, contact information, or other directly identifiable personal information were included. De-identified aggregate statistics, preprocessing details, and experimental scripts may be made available from the corresponding author upon reasonable request and subject to institutional approval. The external validation datasets are publicly available: the Open University Learning Analytics Dataset (OULAD) (Kuzilek et al, 2017) at https://analyse.kmi.open.ac.uk/open_dataset (CC BY 4.0), and EdNet-KT1 (Choi et al, 2020) at https://github.com/riiid/ednet (CC BY-NC 4.0).

Code and Reproducibility Artifacts

The complete pipeline implementation — including the four source-system extractors, stream-merger and manifest-management modules, temporal tensor constructor, LSTM-Attention model (TensorFlow 2.10 and PyTorch 1.13), all baseline models, SHAP attribution, external-validation scripts for OULAD and EdNet-KT1, and a synthetic institutional-data generator — is available for review at https://github.com/Max-YUAN-22/distributed-institutional-risk-inference under the MIT licence. A `reproduce_public.sh` script regenerates all tables and figures derived from publicly available benchmark data; the synthetic generator preserves the feature schema, marginal ranges, temporal structure, and class-imbalance pattern of the institutional dataset and permits independent re-execution of the distributed pipeline and the synthetic scale-out stress test. The trained checkpoint (4.7 MB) is included for inference-only deployment. The private institutional records cannot be publicly released due to institutional data-governance restrictions described under *Ethics Approval and Data Governance*.

Ethics Approval and Data Governance

This study was reviewed and authorised by Dongying Vocational College as the institutional data-governance authority for the secondary research use of anonymised educational administrative data. The authorisation explicitly covers extraction from

the institutional teaching-management system, learning platform, student-affairs records, and school-enterprise cooperation practice records. Before research use, all records were anonymised and de-identified, and no student names, identity-card numbers, contact information, or other directly identifiable personal information were accessed at any stage of the analysis. The data were used solely for educational research and academic analysis under the institution's data-management and academic-ethics requirements, with no commercial purpose and no adverse impact on students' personal rights or interests. Individual informed consent was waived under the institutional policy governing secondary analysis of fully anonymised administrative records, which presents no greater than minimal risk to participants. The institutional authorisation document supporting this declaration is available on request from the corresponding author. Ethical oversight for this study was provided through the institutional data-governance authority cited above; the study did not involve human-subjects intervention or the collection of identifiable private information, and therefore did not fall within the scope of Institutional Review Board review.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgements

The authors thank the Academic Affairs Office and the Information Technology Center of Dongying Vocational College for facilitating access to the anonymised institutional dataset, and acknowledge the Open University (UK) and RIIID for making their datasets publicly available.

Supplementary Files

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

- InstitutionalAuthform.pdf