frontiers in Computer Science

TYPE Original Research
PUBLISHED 20 November 2026
DOI 10.3389/fcomp.2026.1998673

Check for updates

OPEN ACCESS

EDITED BY
Bilal Hassan,
London Metropolitan University,
United Kingdom

REVIEWED BY
Adeel Anjum,
Quaid-i-Azam University, Pakistan
Soumya Prakash Rana,
University of Greenwich,
United Kingdom

*CORRESPONDENCE
Muhammad Adnan Aslam
✉ malikmadnanaslam@gmail.com
Zulfiqar Ali
✉ z.ali@essex.ac.uk

RECEIVED 14 September 2026
REVISED 12 October 2026
ACCEPTED 02 November 2026
PUBLISHED 20 November 2026

CITATION
Aslam MA, Aslam MK, Yasin A and Ali Z (2026) FedPRISM: a privacy-preserving federated learning framework for personalized academic recommendation and early intervention in smart education environments.
Front. Comput. Sci. 8:1998673.
doi: 10.3389/fcomp.2026.1998673

COPYRIGHT
© 2026 Aslam, Aslam, Yasin and Ali. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

FedPRISM: a privacy-preserving federated learning framework for personalized academic recommendation and early intervention in smart education environments

Muhammad Adnan Aslam[1]*, Muhammad Kamran Aslam[2], Amanullah Yasin[3] and Zulfiqar Ali[4]*

[1]Department of Creative Technologies, Faculty of Computing and Artificial Intelligence (FCAI), Air University, Islamabad, Pakistan, [2]Department of Electrical Engineering, National University of Sciences and Technology (NUST) University, Islamabad, Pakistan, [3]Department of Computer Science, Bahria University, Bahria School of Engineering and Applied Sciences (BSEAS), H11, Islamabad, Pakistan, [4]School of Computer Science and Electronic Engineering, University of Essex, Colchester, United Kingdom

Educational risk models are often assessed by aggregate accuracy, although false alarms and leakage in model-selection procedures can materially change their practical value. We present FedPRISM as a reproducible, privacy-aware federated simulation for at-risk screening and preliminary academic guidance, evaluated on SAPEx-D (494 students; 86 encoded predictors). A stratified 80/20 split (395/99) is used; scaling and F-ANOVA selection are fitted within training folds, and a logistic stacking model is trained on five-fold out-of-fold detector scores. The primary input comprises 30 selected non-personality predictors plus all five Big-Five traits (35 dimensions); an ablation uses the 30 non-personality predictors alone. Four approximately equal IID clients are simulated rather than mapped to actual departments. On the held-out set, the stacked model achieves accuracy 0.828, precision 0.825, recall 0.767, F1 = 0.795, false-alarm rate (FAR) = 0.125, and ROC-AUC = 0.889. The Random Forest alone is similar (accuracy 0.818, F1 = 0.775, FAR = 0.107, AUC = 0.894); paired tests do not establish a performance difference. Ablation results indicate that the Random Forest supplies most of the discriminative signal: removing it lowers accuracy to 0.636 and F1 to 0.438. Removing the autoencoder reduces accuracy to 0.816, F1 to 0.781, and ROC-AUC to 0.875, while removing the Isolation Forest yields accuracy of 0.821, F1 of 0.786, and ROC-AUC of 0.881. The ablation analysis did not reveal a statistically significant incremental predictive benefit for the Big-Five personality features examined in the analyzed context, however their contribution might differ depending on the datasets and educational contexts. A separate RDP-accounted Gaussian-noise analysis evaluates client-level privacy for the FedAvg autoencoder component only; it does not cover the forests, stacker, recommendations, or the coordinator's view, and no cryptographic secure aggregation is implemented. The study is a single-institution, retrospective screening experiment: the data do not establish prospective warning time, recommendation uptake, or intervention benefit.

These results support further evaluation, not a claim of deployment readiness or end-to-end differential privacy.

KEYWORDS

at-risk student screening, client-level differential privacy, educational data mining, false-alarm rate, federated-learning simulation, out-of-fold stacking, preliminary guidance, smart education

1 Introduction

Higher education increasingly relies on analysis of student records, learning interactions, and behavioral traces to support planning and student services (George and Wooden, 2023). Such records may combine personal and family circumstances with academic engagement and, in some datasets, psychometric measures such as the Big Five (Aslam et al., 2025). Recent work has studied at-risk ranking, early-warning models, and explainable feedback, but predictive performance alone does not show that an alert arrives early enough or that advice changes learning outcomes (Yoneda et al., 2025; Afzaal et al., 2024; Embarak and Hawarna, 2024). A useful research pipeline must therefore distinguish three questions: whether a model can classify a retrospective risk label, whether the input features precede that label in time, and whether any resulting guidance benefits students. The present study addresses only the first question and prototypes the third; SAPEx-D does not provide the temporal evidence needed to establish the second.

Privacy is an additional design constraint because the SAPEx-D feature set includes family, financial, behavioral, and personality-related information. Federated learning (FL) can reduce the need to pool raw records by coordinating local model training, but FL alone is not a formal privacy guarantee: model updates can still disclose information, and defenses such as differential privacy, secure aggregation, and cryptographic protection have distinct threat models and implementation requirements (Rafi et al., 2024; Li et al., 2024; Kandati and Anusha, 2023). Educational FL studies have begun to evaluate these trade-offs, including cross-course risk ranking and differentially private student-performance prediction (Yoneda et al., 2025; Chen and Qi, 2025). Consequently, claims about privacy must specify the protected unit, the model component, the privacy accountant, and what the aggregation server can observe. We apply this principle explicitly: our experiment is a single-process IID simulation, and its component-level DP analysis is limited to the released autoencoder parameters.

We present FedPRISM as a proof-of-concept research pipeline that combines a four-client federated simulation, heterogeneous risk scores, leakage-controlled stacking, and a rule-based guidance module. The empirical substrate is SAPEx-D, the 494-record dataset described by Aslam et al. (2025), collected at Air University, Islamabad, and spanning personal, family, academic, and Big-Five attributes. The new experiment selects 35 predictors (30 F-ANOVA-ranked non-personality features and the five personality measures), then compares the complete stack with detector and personality ablations. We also report a separate client-level DP sensitivity analysis for the autoencoder updates; the primary ensemble evaluation is unnoised, and the analysis does not claim end-to-end DP or cryptographic secure aggregation. Building on prior SAPEx-D analysis that explored personality-informed performance prediction (Aslam et al., 2024), this work tests whether adding these measures improves held-out risk classification rather than presuming that it does.

The work is guided by the following research questions:

- RQ1. How do a leakage-controlled stacked detector and its individual components perform on the held-out at-risk classification task, including in terms of false alarms?
- RQ2. Does adding the five Big-Five features improve held-out performance relative to the selected 30 non-personality predictors, and which detector contributes most to the stack?
- RQ3. How sensitive is the autoencoder's held-out ranking performance to calibrated client-level Gaussian noise, and what can the current data support about prospective screening and personalized guidance?

The role of personality is treated as an empirical ablation rather than an assumed benefit: we compare otherwise identical 35-feature and 30-feature pipelines on the same outer split and report paired uncertainty. The Good/Average/Poor group profiles are descriptive and are not interpreted as causal or diagnostic effects.

The contributions are fourfold. **(i)** We provide a reproducible four-client IID simulation on SAPEx-D and explicitly delimit it from an operational cross-institution network or cryptographic secure-aggregation deployment. **(ii)** We correct the model-input description and implement training-only scaling/feature selection with five-fold out-of-fold (OOF) score stacking, so the held-out test labels are not used to train the meta-learner or calibrate detector thresholds. **(iii)** We report FAR, ROC-AUC, stratified-bootstrap uncertainty, paired performance comparisons, and detector/personality ablations; these show that the Random Forest is the dominant component and do not establish that the full stack or personality inputs outperform the relevant comparator. **(iv)** We calibrate and report a client-level RDP Gaussian mechanism for the FedAvg autoencoder only, and provide a rule-based guidance prototype whose educational benefit remains to be evaluated.

The remainder of the paper is organized as follows. Section 2 reviews foundational and recent work spanning student-performance prediction, personalized recommendation, and privacy-preserving learning, and articulates the research gaps this work fills. Section 3 details the FedPRISM framework, the SAPEx-D dataset, the preprocessing and feature-selection pipeline, the federated learning and differential-privacy design, the recommendation engine, and the evaluation methodology. Section 4 presents and discusses the experimental results, including a comparison with existing

studies. Section 5 concludes and outlines directions for future work.

2 Related work

This section situates the present work within three intersecting bodies of literature—student-performance prediction, personalized educational recommendation, and privacy-preserving federated learning—covering foundational and recent contributions and the specific gaps that FedPRISM is designed to close.

2.1 Student-performance prediction and educational data mining

Foundational surveys trace the development of educational data mining and student-performance prediction: Romero and Ventura (2007) reviewed work from 1995 to 2005, and Albreiki et al. (2021) later surveyed machine-learning methods for student-performance prediction. Recent educational-data-mining studies have expanded student-performance prediction beyond grades to include behavioral, contextual, and personal factors. Aslam et al. (2024) reported that adding Big-Five measures to SAPEx-D-based academic and demographic predictors increased the coefficient of determination for CGPA prediction from 0.63 to 0.83, while using SHAP to inspect feature contributions. Their subsequent dataset descriptor documents the 494-record SAPEx-D cohort and its personal, family, academic, and personality fields (Aslam et al., 2025). Across other settings, Zhao et al. (2023) compared machine-learning models for academic performance, while Huang and Chen (2024) modeled temporal interactions in MOOCs; these studies illustrate both the expansion of predictor sets and the dependence of reported performance on dataset and evaluation design.

Complementary work examines institutional data integration and warning systems. Fahd and Miah (2023) evaluated a big-data analytics approach for student-success factors, and Hanaysha et al. (2023) analyzed classroom and institutional conditions associated with engagement and performance. Khan et al. (2021) examined AI-based student-performance monitoring and preventive measures. Embarak and Hawarna (2024) introduced RADAR, an explainable framework for at-risk monitoring. Al-Tameemi et al. (2024) examined a hybrid machine-learning approach for multi-class student-performance prediction. López-García et al. (2023) used XGBoost with TOPSIS-based feature extraction and ADASYN oversampling for early detection of students' academic failure. Pawar and Dongardive (2024) examined feature-selection effects using assessment data. These studies reinforce that feature definitions, class balance, cohort composition, and temporal availability all affect what an apparent early-warning result means.

Notwithstanding this progress, prediction is often evaluated separately from the subsequent student-support process. The present work therefore prototypes a narrow bridge from a risk score to candidate behavior-focused prompts for human review. These prompts are not validated recommendations, and no behavioral change or academic outcome is claimed.

2.2 Personalized recommendation systems in education

Educational recommendation research spans course selection, learning-resource ranking, and personalized feedback. Anupama and Elayidom (2022) reviewed existing course-recommendation methods spanning collaborative filtering, machine learning, and topic modeling, while Zhong (2022) proposed a student-management recommender based on collaborative filtering. Ceyhan et al. (2021) also studied collaborative filtering for course recommendation, and Embarak (2021) explored a machine-learning recommendation system for adaptive education. Zhalgassova et al. (2023) explored personality-informed recommendations for extracurricular activities. Afzaal et al. (2024) first described automatic recommendations to support student self-regulation (Afzaal et al., 2021); later work examined explainable, data-driven feedback. Takami et al. (2023) studied personality-tailored explanations in a quiz recommender; their 18-day study reported increased use of recommendations in the tailored group. These findings motivate personalized feedback as a research direction, but effects from those evaluated platforms cannot be transferred to the rule-based guidance prototype in the present study.

This literature shows that educational recommenders can expose learners to tailored resources and feedback, and that some studies have evaluated learner engagement. It does not establish that every personalization strategy improves academic outcomes. Personality and behavioral variables can support personalization, but they are sensitive and require careful treatment; moreover, the present study's guidance module is not user-tested and is not an evaluated educational intervention.

2.3 Privacy-preserving and federated learning in education

The privacy of educational data has risen to prominence as the sensitivity of learning analytics has become undeniable. Federated learning, which enables collaborative model training without exchanging raw data, is now the subject of extensive methodological study in the broader machine-learning literature. Rafi et al. (2024) provided a comprehensive survey of fairness and privacy preservation in federated learning, cataloging privacy attacks, defenses, and the trade-offs they induce. Li et al. (2024) surveyed federated-learning data security and privacy-preserving techniques in edge–Internet-of-Things settings, while Kandati and Anusha (2023) examined the security and privacy challenges of federated systems, including differential privacy and homomorphic encryption as protective mechanisms. These studies emphasize that the privacy claim must be matched to the mechanism, protected unit, and observer model.

Within education, recent federated studies include Yoneda et al. (2025), who evaluated ranking-based at-risk prediction with federated learning and differential features across courses, and Chen and Qi (2025), who proposed entropy-adaptive differentially private FedAvg for programming-course performance prediction. These studies provide relevant evidence that distributed

training and privacy mechanisms are being explored in educational settings, but their datasets, model designs, privacy definitions, and evaluation protocols differ from SAPEx-D and from one another. Moshawrab et al. (2023) review federated aggregation algorithms and their limitations; this is a general FL survey, not an educational prediction experiment. Taken together, this literature motivates careful component-level evaluation. The present work adds an explicit, leakage-controlled screening experiment and guidance prototype, while limiting its privacy conclusions to the mechanism and threat model actually implemented.

2.4 Research gaps and positioning

Three practical gaps motivate this study. First, prediction, personalized feedback, and privacy-aware training are often evaluated separately; there is value in a reproducible pipeline that exposes how these components interact, without presuming that their combination improves outcomes. Second, SAPEx-D provides a setting in which personal, family, academic, and Big-Five predictors can be examined together, but its single-institution cross-sectional design requires conservative interpretation. Third, privacy claims in federated systems depend on the precise adjacency relation, aggregation mechanism, and observer model. We therefore report a scoped client-level DP sensitivity analysis for the autoencoder and make explicit that the complete ensemble is not end-to-end differentially private.

3 Proposed methodology

3.1 Framework overview

FedPRISM is evaluated as an in-process federated-learning simulation that maps the 395-record outer training partition to four approximately balanced IID shards. It converts heterogeneous predictors into (i) a binary retrospective risk classification and (ii) rule-based preliminary guidance. Figure 1 shows the implemented workflow, including training-only reduction of 86 encoded candidates to 35 base-model inputs (30 non-personality predictors plus five Big-Five traits; 30 inputs in the no-Big-Five ablation), client-local fitting, FedAvg updates, averaging of client-model scores, OOF-trained stacking, and guidance generation. The stacker receives three detector scores. This is not an operational network: the shards do not correspond to observed departments, no network transport or cryptographic secure aggregation is implemented, and the primary ensemble results use unnoised updates. Only the separate autoencoder experiment adds central Gaussian noise to clipped client updates.

3.2 SAPEx-D dataset

The study is conducted on the SAPEx-D (Student Academic Performance Exploration) dataset (Aslam et al., 2025), collected at Air University, Islamabad, Pakistan, between September 2022 and April 2024. SAPEx-D comprises 494 student records spanning four factor families: *personal* factors (age, gender, transportation, accommodation, extracurricular engagement, sleep, part-time work), *family* factors (parental education and occupation, income, parental status and involvement, number of siblings), *academic* factors (degree major, attendance, study hours, reading frequency, scholarship type, note-taking, class engagement, exam preparation, and assessment performance), and *personality* factors derived from the Big-Five inventory (openness, conscientiousness, extraversion, agreeableness, neuroticism). The dataset is notable for combining all four factor families in a single student-level schema; its descriptor states that the data are available on reasonable request rather than as an unrestricted public download (Aslam et al., 2025).

Following the institutional grading policy, students are categorized into three performance bands from their CGPA: *Good* ($3.67 \leq$ CGPA ≤ 4.00), *Average* ($2.67 \leq$ CGPA < 3.67), and *Poor* ($1.00 \leq$ CGPA < 2.67). The resulting distribution—37 Good, 241 Average, and 216 Poor students—is depicted in Figure 2. For the at-risk detection task, the *Poor* band constitutes the positive (anomaly) class, while Good and Average students form the negative (on-track) class.

Figure 3 presents the continuous CGPA distribution with the three policy-derived bands superimposed, revealing a left-skewed profile concentrated around the 3.0 mark—consistent with the cohort composition reported in the dataset descriptor (Aslam et al., 2025)—and a substantial at-risk tail extending toward the 1.6 floor. Figure 4 summarizes the demographic composition: 343 male (69.4%) and 151 female (30.6%) participants, with 434 students aged 18–21 and 60 aged 22–25, and a letter-grade distribution spanning the eight levels of the institutional scale.

3.3 Data preprocessing

The supplied SAPEx-D file contains 494 complete records, 86 encoded predictors, and two outcome fields (Exact CGPA and Letter Grade) (Aslam et al., 2025). We remove the outcome columns from the predictor matrix and derive the three-band and binary labels from Exact CGPA. Binary one-hot indicators are left as 0/1; ordinal codes and continuous personality variables are min–max scaled. Crucially, the scaler is fitted only on the current training partition and then applied unchanged to its validation or test partition:

$$X_j^* = \frac{X_j - \min(X_j^{\text{train}})}{\max(X_j^{\text{train}}) - \min(X_j^{\text{train}})}, \tag{1}$$

where the extrema are learned from the training rows only. This operation is repeated inside each meta-training fold and once for the outer train/test split. The held-out test observations therefore contribute neither to scaling nor to feature selection, threshold calibration, or meta-learner fitting. No imputation is required for this complete release.

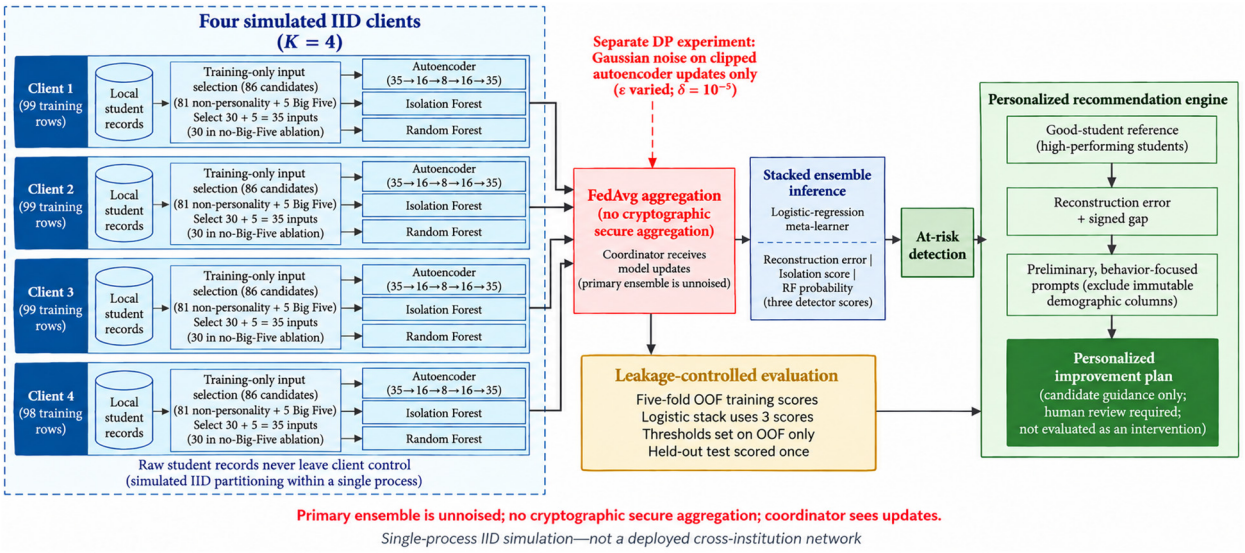

FIGURE 1
Implemented FedPRISM workflow. From 86 encoded candidates, training-only selection yields 35 inputs per base detector (30 non-personality predictors plus five Big-Five traits; the ablation uses 30); the logistic stacker receives three detector scores. The four IID clients are simulated in one process. Gaussian noise and client-level RDP accounting are applied only in the separate autoencoder sensitivity experiment. No cryptographic secure aggregation or end-to-end privacy guarantee is implemented.

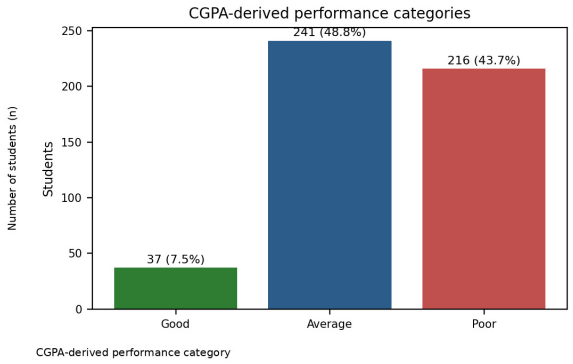

FIGURE 2
Distribution of the 494 SAPEx-D students across the three performance categories derived from the CGPA grading policy.

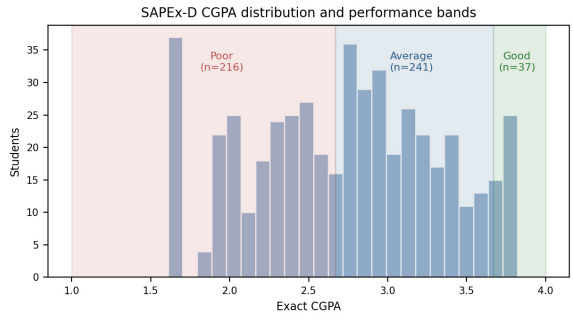

FIGURE 3
Distribution of Exact CGPA with the Good, Average, and Poor performance bands highlighted. The at-risk (Poor) tail is clearly delineated on the left of the distribution.

3.4 Feature extraction via F-ANOVA

For the retrospective screening task, candidate predictors are ranked by univariate F-statistics against continuous CGPA. This ranking is used for base-model input selection only; the guidance module separately whitelists modifiable academic-behavior features and does not convert demographic or personality rankings into advice. For each candidate feature, the F-statistic from an F-test for regression against CGPA is computed:

$$F = \frac{\text{explained variance}/1}{\text{residual variance}/(n-2)}. \quad (2)$$

The input release has 86 encoded predictors, of which five are Big-Five traits and 81 are non-personality candidates; these are the selection pool, not the final model input dimensionality. On each training partition, we retain the 30 highest-scoring non-personality predictors from the personal, family, academic, and behavioral/contextual fields, then append all five Big-Five traits, producing a 35-dimensional vector. The autoencoder, Isolation Forest, and Random Forest each receive the selected 35-dimensional input; the logistic stack receives the three resulting detector scores. In the personality ablation, the base detectors receive 30 dimensions (and the stack still receives three scores). In each OOF fold, scaling and F-ANOVA ranking are refitted using only that fold's training rows. In the outer model, feature selection is fitted on the 395-record training partition and applied to the untouched 99-record test partition. The training-set ranking in Figure 5 places gender indicators, major, scholarship, and several contextual variables near the top; extraversion has the highest F-score among the five personality traits. F-scores are univariate associations and do not

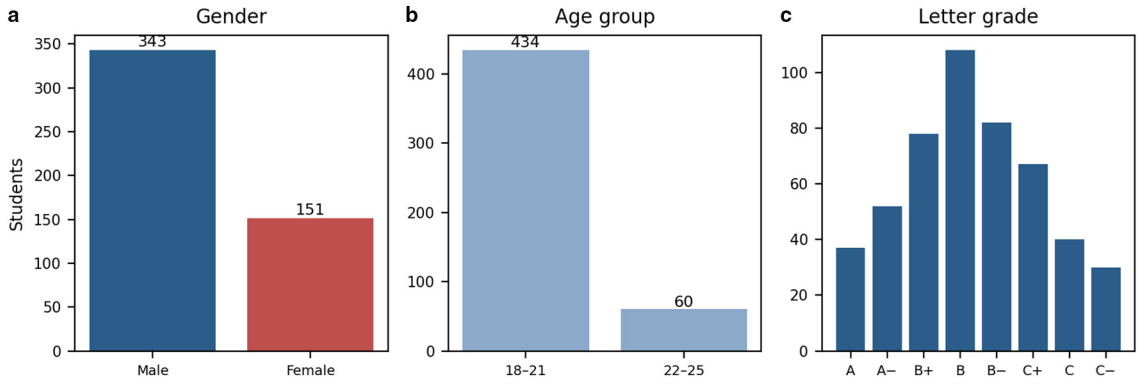

FIGURE 4
Demographic composition of SAPEx-D: **(a)** gender, **(b)** age group, and **(c)** letter-grade distribution across the eight institutional grade levels.

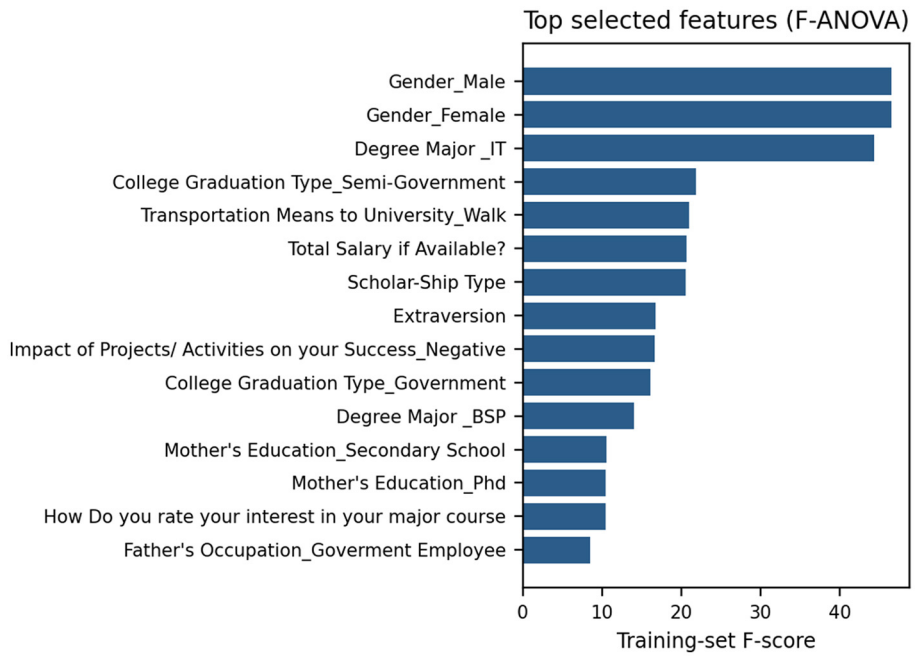

FIGURE 5
Top 15 training-set F-ANOVA scores among selected predictors. Gender indicators and contextual variables rank highly; extraversion has the highest *F*-score among the Big-Five traits. Scores are univariate and are not measures of causality or ablation benefit.

establish causal effects or incremental predictive value beyond the ablations.

3.5 Federated simulation and component-level privacy

3.5.1 Client partitioning

The outer training set contains 395 students and is randomly permuted into four approximately equal IID shards using a fixed seed. The resulting sizes and binary-label counts are: client 1, 99 (56 on-track, 43 at-risk); client 2, 99 (54, 45); client 3, 99 (55, 44); and client 4, 98 (57, 41). These are simulated clients, not observed department-level or institutional partitions; the split is not a non-IID experiment. Figure 6 shows the final outer-training distribution. The same deterministic partitioning rule is applied within each OOF fold.

3.5.2 Local detectors and score fusion

The autoencoder has architecture $35 \to 16 \to 8 \to 16 \to 35$ and is trained on the on-track (Good and Average) examples. At each of five FedAvg rounds, each client initializes from the current global weights and performs 10 full-batch local epochs. An Isolation Forest and a class-balanced Random Forest are fitted separately on each simulated client's training records. For the experiment, each

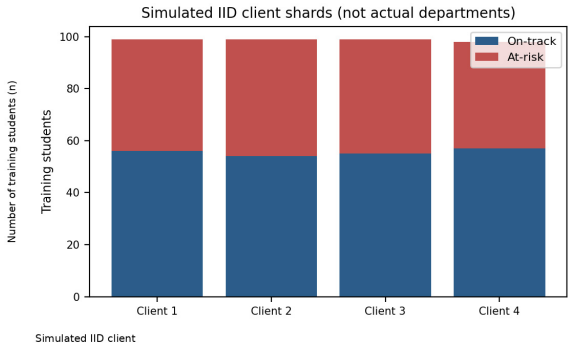

FIGURE 6
Label composition of the four approximately balanced IID training shards. These are simulated client partitions, not actual departments or independently collected institutional datasets.

test observation is scored by all four client forests: the Isolation-Forest score is the arithmetic mean of the negative client decision functions, and the Random-Forest score is the arithmetic mean of client probabilities for the at-risk class. The autoencoder supplies one reconstruction-error score from the FedAvg model. These three scores form the inputs to logistic stacking. In the current single-process simulation, score averaging occurs centrally; the client forest parameters and per-student scores are not protected by secure aggregation or a DP mechanism.

3.5.3 Leakage-controlled stacking

To avoid fitting the meta-learner on in-sample base-model predictions, detector scores for the outer training set are generated with five-fold stratified cross-fitting. For each fold, scaling, F-ANOVA selection, and the client models are fitted only on the other four folds, and scores are recorded for the held-out fold. The logistic-regression meta-learner (with standardization) is fitted on the resulting OOF score matrix and training labels. Base-detector thresholds are selected by maximizing Youden's J on those same OOF training scores. Final base models are then fitted on all 395 outer-training examples; the 99 outer-test examples are scored once, and neither test labels nor test scores are used in model selection, threshold calibration, or meta-training. The centralized comparator follows the same outer split, fold-specific preprocessing, and OOF protocol.

3.5.4 Differential-privacy mechanism and scope

Following the approximate differential-privacy framework formalized by Dwork and Roth (2014), a separate sensitivity experiment applies central Gaussian noise to the clipped, uniformly averaged client updates of the autoencoder. Each full client update is clipped to L2 norm $C = 1.0$. Under replace-one-client adjacency (the protected unit is a simulated shard, not an individual student) and $K = 4$ fixed participants, the sensitivity of the mean update is bounded by $\Delta = 2C/K = 0.5$. For an overall privacy target (ε, δ), Gaussian noise with standard deviation $z\Delta$ is added

TABLE 1 Client-level DP calibration and held-out autoencoder utility.

Overall ε	Noise multiplier z	Update noise SD $z\Delta$	AE AUC	AUC 95% CI
No noise (reference)	–	0.000	0.618	[0.507, 0.721]
0.5	22.457	11.229	0.640	[0.524, 0.750]
1	11.264	5.632	0.440	[0.318, 0.560]
2	5.697	2.848	0.456	[0.341, 0.570]
5	2.427	1.214	0.571	[0.459, 0.683]
10	1.274	0.637	0.551	[0.441, 0.668]

The no-noise reference is not a DP setting. The noise standard deviation is per coordinate of the averaged update.

to every aggregate update in each of the $T = 5$ rounds. The noise multiplier z is calibrated using a Renyi differential privacy (RDP) accountant with no client-subsampling amplification and basic composition across rounds. The mechanism is implemented in `federated.py`; the sweep runs one seeded model/noise realization per target ε:

$$\varepsilon(\delta) = \min_{\alpha \in \mathcal{A}} \left[\frac{T\alpha}{2z^2} + \frac{\ln(1/\delta)}{\alpha - 1} \right], \quad \mathcal{A} = \{1.25, 1.5, 1.75, 2, 3, 4, 5, 8, 10, 16, 20, 32, 64, 128, 256, 512, 1024\}. \tag{3}$$

The sweep uses $\delta = 10^{-5}$ and target $\varepsilon \in \{0.5, 1, 2, 5, 10\}$; the calibrated noise multipliers and aggregate standard deviations are reported in Table 1. This is client-level approximate DP for release of the autoencoder parameters only. It does not cover the RF/IF models, the stacking learner, recommendation output, or an untrusted coordinator's view of the individual updates. The coordinator in this code can inspect those updates before adding noise; no cryptographic secure-aggregation protocol is implemented. Moreover, the primary ensemble results below use no added noise. Accordingly, they are not end-to-end differentially private results.

Figure 7 reports the final unnoised model's in-simulation training reconstruction loss after each FedAvg round. This diagnostic is computed on the aggregated on-track training records inside the experiment process and is not evidence of cross-site convergence or privacy protection.

3.6 Personalized recommendation engine

The final stage is a rule-based guidance prototype that translates an at-risk flag into candidate prompts for human review, not a validated recommender or intervention. When supplied with a flagged student's feature vector, it compares eligible actionable attributes with the mean of the *Good*-band records in the training partition and computes the autoencoder's absolute feature-wise reconstruction error. For each actionable attribute, it also computes the signed gap (Good-band mean minus the student's value).

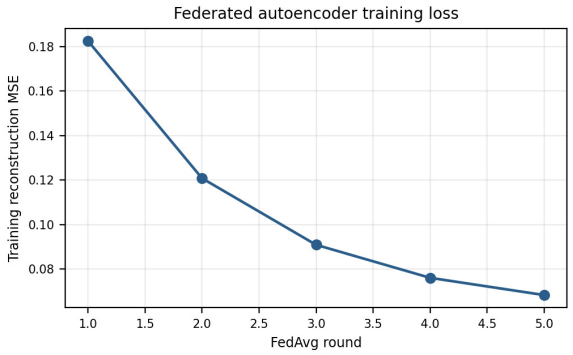

FIGURE 7
Training reconstruction MSE of the unnoised FedAvg autoencoder across five communication rounds, evaluated in the single-process simulation. The curve is a training diagnostic, not a held-out generalization or privacy result.

The prototype assigns higher review priority when a feature's reconstruction error is at or above the 70th percentile of that student's feature-wise errors or when the signed gap is positive; prompts are sorted by review priority and gap. A curated mapping supplies candidate discussion prompts concerning, for example, attendance barriers, weekly study planning, note-taking, reading, course interest, or seeking information about financial-aid options.

The current mapping is limited to modifiable academic-behavior attributes and support-seeking topics; it excludes Big-Five personality traits and immutable demographic/context fields. It does not implement separate *improvement* and *reinforcement* classes, infer causal deficiencies, or establish that a suggested behavior will change performance. This standalone prototype is separate from the detector-performance evaluation. Its prompts are preliminary associations against a single-institution reference group; they have not been evaluated with students or advisers, and no student-facing deployment, uptake, warning lead time, or learning-outcome effect is claimed.

3.7 Evaluation methodology

The primary evaluation uses a stratified 80/20 split with random seed 42 (395 training and 99 held-out test records; 43 positive and 56 negative test cases). We report accuracy, precision, recall, F1, false-alarm rate, and ROC-AUC. FAR is important because it is the proportion of on-track students incorrectly flagged as at-risk :

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \text{Precision} = \frac{TP}{TP + FP}, \quad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN}, F1 = \frac{2\,\text{Precision}\,\text{Recall}}{\text{Precision} + \text{Recall}}, \quad (5)$$

$$\text{FAR} = \frac{FP}{FP + TN}. \quad (6)$$

Here, the positive class is Poor (at-risk), and the negative class is Good or Average (on-track). ROC-AUC is computed from continuous detector scores. For each point estimate we calculate a 95% percentile interval from 2,000 class-stratified bootstrap resamples of the 99 held-out observations, preserving the test-set counts of 43 positive and 56 negative records. These intervals quantify held-out-sample uncertainty conditional on the fitted models; they do not include retraining, institution-to-institution, or temporal variation. Paired stratified-bootstrap intervals are calculated for differences in accuracy, F1, FAR, and ROC-AUC. Exact two-sided McNemar tests compare paired correctness for the complete stack against the Random Forest, centralized comparator, and 30-feature no-personality model; comparisons with the weak unsupervised detectors are treated as secondary.

Detector thresholds are selected by maximizing Youden's J on five-fold OOF training scores only. The logistic stack is fitted to those OOF detector scores, with score standardization learned on the OOF training matrix. The centralized comparator uses the same outer split, fold-wise preprocessing, OOF stacking, and threshold procedure. The primary model comparison is unnoised. The differential-privacy analysis is a separate autoencoder-only sensitivity experiment, with the accountant and mechanism defined in Section 3; it is not a privacy guarantee for the full ensemble.

4 Results and discussion

4.1 Feature-importance analysis

The training-set F-ANOVA ranking (Figure 5) places the gender indicators (each $F = 46.49$), the IT-major indicator ($F = 44.39$), and several college, travel, financial, and scholarship categories among the largest univariate scores. Extraversion has the largest F-score among the five personality variables ($F = 16.81$); the remaining Big-Five scores are substantially smaller (all $F \leq 1.64$). These rankings are descriptive, and the apparent prominence of a one-hot category does not imply a causal effect. The signed full-cohort correlations in Figure 8 likewise are unadjusted: scholarship-type and major-interest codes correlate positively with CGPA ($r = 0.201$ and $r = 0.171$), class attendance has a smaller positive association ($r = 0.118$), while the encoded salary category is negatively associated ($r = -0.224$) and extraversion has $r = -0.188$. Because several variables are ordinal category codes, the signs and magnitudes should not be interpreted as causal or interval-scale effects. These full-cohort descriptive plots were not used for model fitting, feature selection, or threshold calibration.

Figure 9 displays descriptive Big-Five means for Good and Poor students. Conscientiousness is higher in the Good group (69.8 vs. 59.6), whereas neuroticism is lower (48.5 vs. 53.2); the group means for the other traits are closer. This pattern motivates testing the traits but does not show that they add predictive value to the classifier. Indeed, the held-out personality ablation below provides no statistically clear evidence of improvement from including the five traits.

4.2 At-risk detection performance

Table 2 reports point estimates on the held-out test set. The stacked model classifies 33 of 43 at-risk students correctly (recall 0.767) and raises 7 false alarms among 56 on-track students (FAR

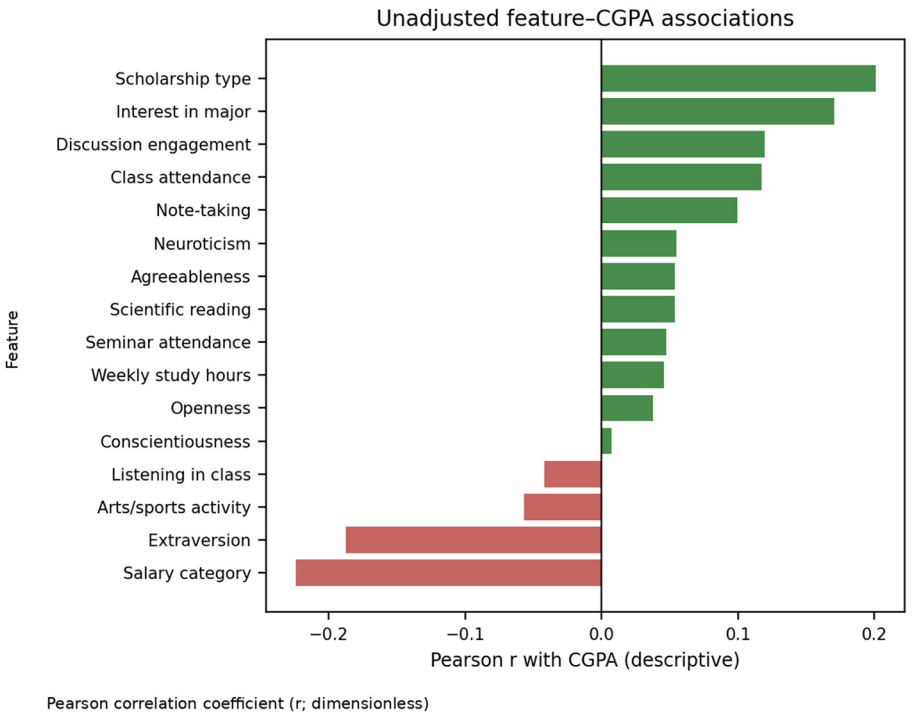

FIGURE 8
Unadjusted Pearson correlations between selected encoded variables and CGPA. These are descriptive associations; in particular, correlations with ordinal category codes depend on the chosen coding and are not causal effects.

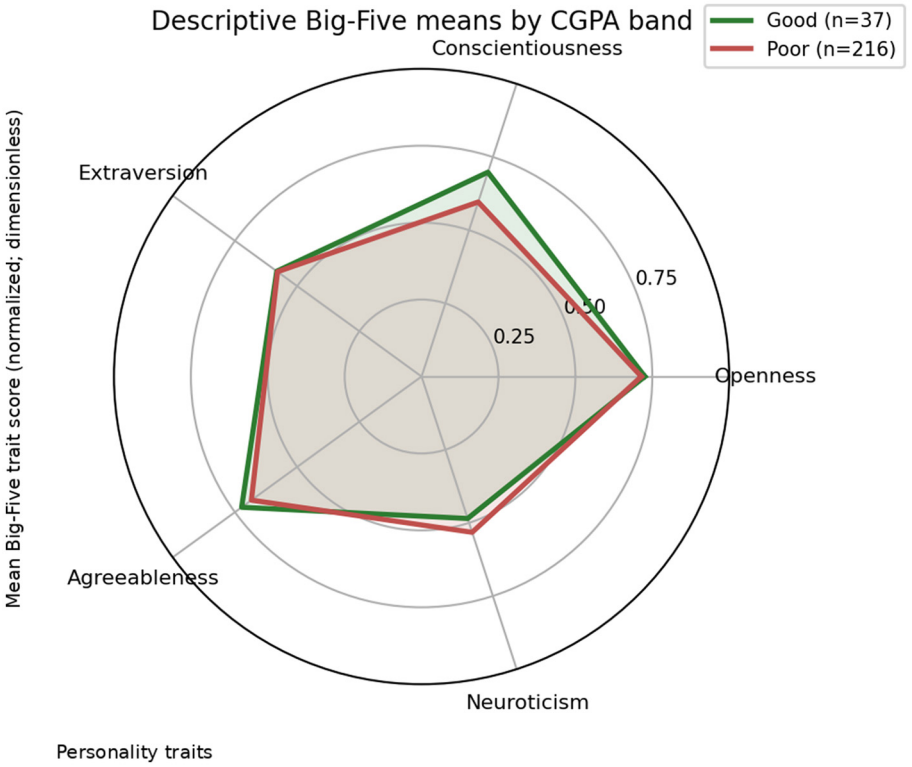

FIGURE 9
Descriptive Big-Five mean profiles for Good and Poor students, shown on the 0–1 scale. Group differences are not causal and are not used as personality-change prescriptions.

TABLE 2 Held-out performance on the 99-student test partition.

Model	Accuracy	Precision	Recall	F1	FAR	ROC-AUC
Autoencoder	0.596	0.535	0.535	0.535	0.357	0.580
Isolation Forest	0.444	0.423	0.767	0.545	0.804	0.517
Random Forest	0.818	0.838	0.721	0.775	0.107	0.894
FedPRISM (full, 35 inputs)	**0.828**	**0.825**	**0.767**	**0.795**	**0.125**	**0.889**
Centralized FedPRISM comparator	0.838	0.814	0.814	0.814	0.143	0.922

The first four rows are the simulated federated models; the final row is the centralized FedPRISM comparator. FAR is $FP/(FP + TN)$; at-risk is the positive class. Bold values indicate the performance metrics of the proposed FedPRISM model (full configuration with 35 inputs).

TABLE 3 Stratified-bootstrap 95% confidence intervals for held-out performance (2,000 resamples; same fixed fitted models), including the centralized comparator.

Model	Accuracy	Precision	Recall	F1	FAR	ROC-AUC
Autoencoder	[0.495, 0.687]	[0.424, 0.650]	[0.372, 0.674]	[0.410, 0.644]	[0.232, 0.482]	[0.462, 0.691]
Isolation Forest	[0.364, 0.525]	[0.371, 0.474]	[0.628, 0.884]	[0.475, 0.612]	[0.696, 0.911]	[0.399, 0.634]
Random Forest	[0.737, 0.889]	[0.732, 0.943]	[0.581, 0.837]	[0.667, 0.867]	[0.036, 0.196]	[0.822, 0.949]
FedPRISM (full, 35 inputs)	[0.747, 0.899]	[0.721, 0.938]	[0.628, 0.884]	[0.699, 0.881]	[0.036, 0.214]	[0.815, 0.947]
Centralized FedPRISM comparator	[0.758, 0.909]	[0.717, 0.917]	[0.698, 0.930]	[0.723, 0.892]	[0.054, 0.232]	[0.860, 0.966]

$= 0.125$). The Random Forest classifies 31 at-risk students correctly and raises 6 false alarms (FAR $= 0.107$). Thus, the stack gains two true positives at the cost of one additional false positive on this test split; its accuracy (0.828) and F1 (0.795) are only slightly higher than the Random Forest values (0.818 and 0.775), while its ROC-AUC is slightly lower (0.889 vs. 0.894). The autoencoder and Isolation Forest alone perform poorly at the selected operating thresholds, especially for false alarms (FAR $= 0.357$ and 0.804, respectively). These results do not support the claim that the stack clearly outperforms its strongest constituent.

Table 3 gives class-stratified percentile-bootstrap 95% intervals over the test cases. These intervals are wide for all models because the held-out sample contains only 99 students; they are conditional on the single fitted split and should not be read as evidence of generalization across institutions.

Figure 10 shows the held-out error counts, while Figure 11 summarizes point estimates across models; uncertainty should be interpreted from Table 3 and the paired tests below. The centralized FedPRISM comparator reached accuracy 0.838, F1 $= 0.814$, FAR $=$ 0.143, and ROC-AUC $= 0.922$ (Table 2). Paired-bootstrap intervals for the federated-stack-minus-centralized differences were [−0.061, 0.040] for accuracy, [−0.094, 0.049] for F1, [−0.054, 0.000] for FAR, and [−0.068, 0.000] for AUC; the exact McNemar test was $p = 1.00$ with seven discordant cases. The results do not resolve a difference on this small fixed test set and should not be interpreted as proof of equivalence.

The held-out reconstruction-error distributions in Figure 12 overlap substantially; the median is 0.067 for on-track and 0.079 for at-risk students. This modest separation is consistent with the autoencoder's limited standalone ranking performance (AUC $=$ 0.580), rather than evidence of a strong anomaly signal. Figure 13 shows the Random Forest AUC at 0.894 and the full stack at 0.889, compared with 0.580 for the autoencoder and 0.517 for Isolation Forest. The FAR plot (Figure 14) reinforces the practical distinction: the ensemble yields seven false positives among 56 on-track students, compared with six for the Random Forest; the unsupervised Isolation Forest yields 45. At this operating point, seven of 56 on-track students would trigger an unnecessary review in the test sample; in practice, false alerts can consume limited advising capacity and may stigmatize students if treated as labels. Thresholds therefore require locally governed trade-offs between missed at-risk students and false alarms, followed by human review; this single test split cannot identify a universally appropriate operating point.

4.3 False alarm rate (FAR) analysis

Table 4 gives a dedicated count-based comparison of false alarms on the same held-out SAPEx-D test set. FAR is $FP/(FP + TN)$, so the denominator is the 56 on-track cases; it is not the proportion of all alerts that are false. The full stack produced seven false alerts ($7/56 = 0.125$), compared with six for the Random Forest, eight for the centralized comparator, and nine for the no-Big-Five stack. With only 56 negative cases, one additional false alert changes FAR by approximately 0.018. These point estimates therefore describe this split, not a deployment operating point. In any future evaluation, threshold selection should be governed locally in light of advising capacity and the relative costs of missed cases and unnecessary follow-up, and flagged records should remain subject to human review.

4.4 Ablation study and paired comparisons

Table 5 isolates the detector and feature contributions using the same held-out observations; Table 6 reports paired-bootstrap intervals and exact McNemar comparisons for the principal contrasts. Removing the Random Forest from the stack reduces accuracy from 0.828 to 0.636 and F1 from 0.795 to 0.438 (ROC-AUC falls from 0.889 to 0.674). The updated ablation results also show lower performance when either the autoencoder or Isolation

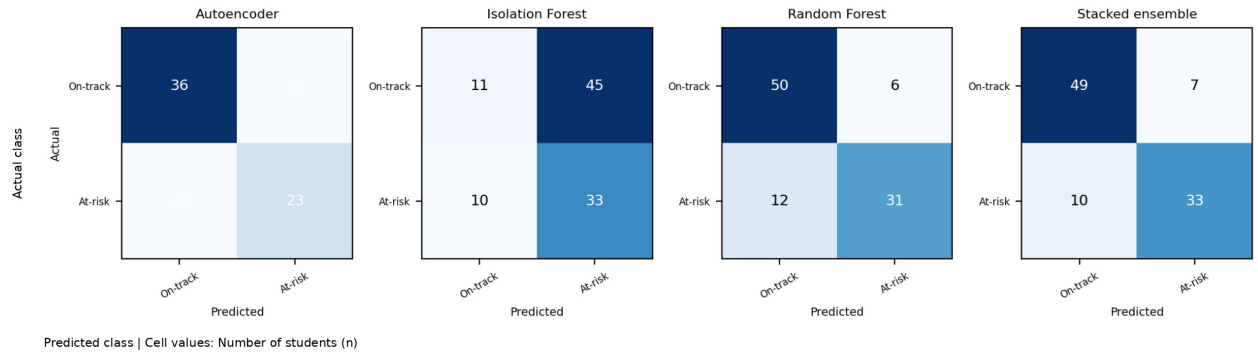

FIGURE 10
Confusion matrices on the 99-record held-out test set. Cells show counts for on-track (negative) and at-risk (positive) students; FAR is computed from the first row's off-diagonal count relative to all on-track cases.

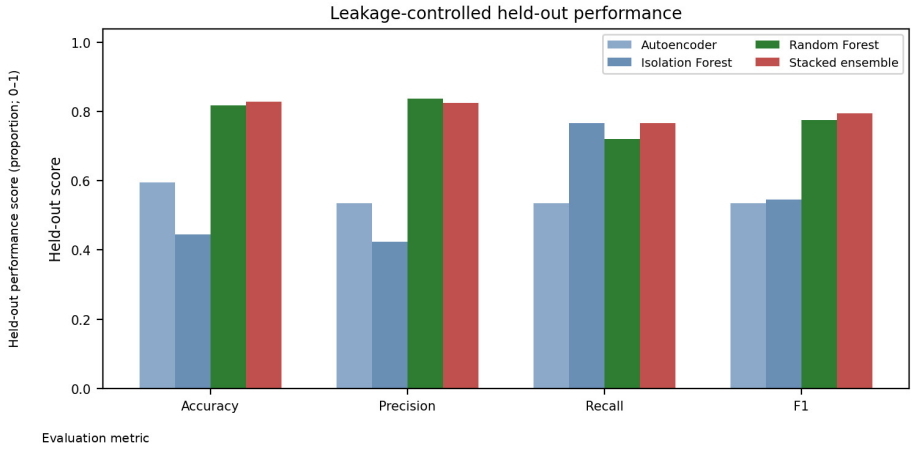

FIGURE 11
Point-estimate comparison of accuracy, precision, recall, and F1 on the held-out set. Bootstrap intervals are reported in Table 3.

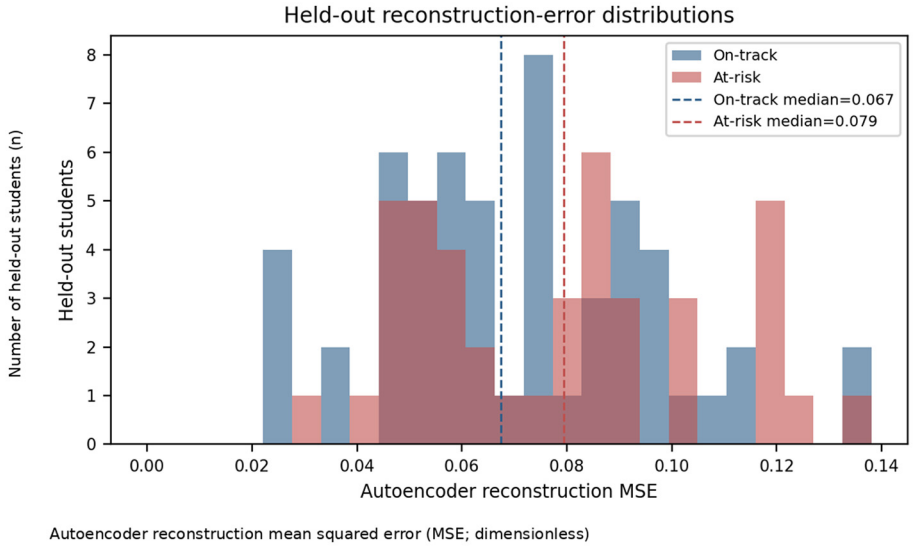

FIGURE 12
Held-out reconstruction-error distributions for on-track and at-risk students. Medians (0.067 and 0.079, respectively) differ modestly, with substantial overlap; this is a weak anomaly signal rather than decisive separation.

Forest is removed. Without the autoencoder, accuracy decreases to 0.816, F1 to 0.781, and ROC-AUC to 0.875. Without the Isolation Forest, accuracy decreases to 0.821, F1 to 0.786, and ROC-AUC to 0.881. These results suggest that both components may contribute to the predictive performance of the full framework, although the magnitude and statistical reliability of their contributions require further paired evaluation. The equal-weight standardized blend is weaker than the learned stack, but this does not establish that stacking improves on the Random Forest: the full model's point estimates are only +0.010 in accuracy and +0.020 in F1, its ROC-AUC is 0.005 lower, and its FAR is 0.018 higher. The paired-bootstrap 95% intervals for full-stack minus Random Forest accuracy [(−0.020, 0.051)], F1 [(−0.020, 0.073)], FAR [(0.000, 0.054)], and ROC-AUC [(−0.018, 0.010)] include zero; the exact McNemar test is not significant ($p = 1.00$; only three discordant cases).

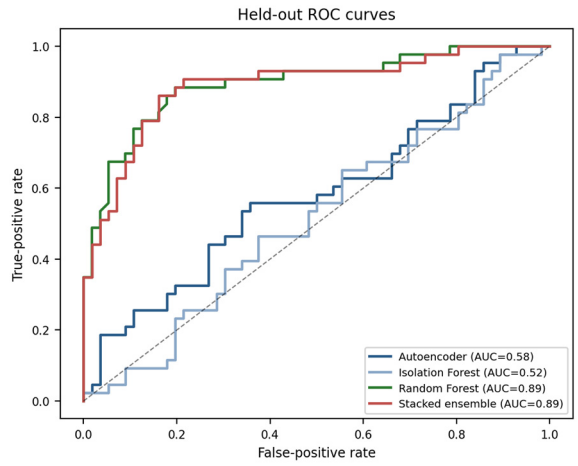

FIGURE 13
Held-out ROC curves. Random Forest has AUC 0.894 and the stacked model 0.889; the unsupervised autoencoder and Isolation Forest have AUC 0.580 and 0.517, respectively.

The 30-feature no-personality stack has the same accuracy as the 35-feature stack (0.828), slightly higher F1 (0.805 vs. 0.795), and lower AUC (0.861 vs. 0.889). However, paired-bootstrap intervals for the full-minus-no-personality accuracy difference [(−0.061, 0.061)], F1 difference [(−0.091, 0.062)], FAR difference [(−0.107, 0.036)], and AUC difference [(−0.002, 0.064)] all include zero; McNemar's exact test is $p = 1.00$. Accordingly, these data do not establish an incremental classification benefit from the five personality measures. This task-specific result is distinct from the earlier CGPA-regression comparison by Aslam et al. (2024), which reported an increase in R^2 from 0.63 to 0.83 after adding Big-Five variables; that prior result does not establish incremental

TABLE 4 False-alarm comparison for base detectors, ensemble variants, and the centralized comparator on the held-out SAPEx-D test set.

Dataset	Model	FP	TN	FAR
SAPEx-D	Autoencoder	20	36	0.357
SAPEx-D	Isolation Forest	45	11	0.804
SAPEx-D	Random Forest	6	50	0.107
SAPEx-D	FedPRISM stack (35 inputs)	7	49	0.125
SAPEx-D	FedPRISM without autoencoder	7	49	0.125
SAPEx-D	FedPRISM without Isolation Forest	7	49	0.125
SAPEx-D	FedPRISM without Random Forest	7	49	0.125
SAPEx-D	FedPRISM equal-weight score blend	15	41	0.268
SAPEx-D	Centralized FedPRISM comparator	8	48	0.143
SAPEx-D	FedPRISM without Big Five (30 inputs)	9	47	0.161

FP denotes on-track students incorrectly flagged as at-risk; TN denotes correctly classified on-track students. FAR is $FP/(FP + TN)$, with 56 on-track cases in every row.

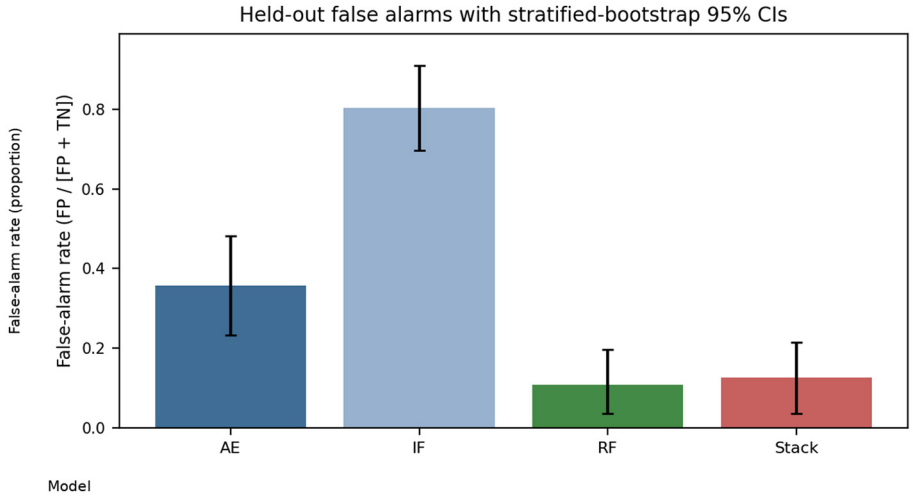

FIGURE 14
False-alarm rates on the 99-case test set with class-stratified-bootstrap 95% intervals. FAR is the fraction of on-track students flagged at risk.

TABLE 5 Detector and feature ablations on the held-out test set.

Configuration	Accuracy	F1	ROC-AUC
Autoencoder only	0.596	0.535	0.580
Isolation Forest only	0.444	0.545	0.517
Random Forest only	0.818	0.775	0.894
FedPRISM (full, 35 features)	0.828	0.795	0.889
FedPRISM without autoencoder	0.816	0.781	0.875
FedPRISM without Isolation Forest	0.821	0.786	0.881
FedPRISM without Random Forest	0.636	0.438	0.674
FedPRISM equal-weight score blend	0.616	0.513	0.722
FedPRISM without Big Five (30 features)	0.828	0.805	0.861

The full and no-Big-Five stacks use separately cross-fitted meta-learners.

benefit for the present retrospective binary at-risk classification. Figure 15 visualizes the principal point estimates; the confidence intervals and paired tests, rather than rank ordering alone, should guide interpretation.

4.5 Autoencoder-level differential-privacy sensitivity

Table 1 and Figure 16 report the separate client-level DP analysis for the autoencoder. The table gives the calibrated Gaussian noise multiplier and standard deviation on the aggregated update for each target overall ε; $\delta = 10^{-5}$, five communication rounds, four clients, and update clipping norm $C = 1$ are fixed. Because there are only four simulated clients, the calibrated noise is large at the smaller privacy budgets. Each ε is represented by one seeded model/noise run. The held-out AUCs are 0.640, 0.440, 0.456, 0.571, and 0.551 for $\varepsilon = 0.5, 1, 2, 5, 10$, respectively, and vary non-monotonically. The no-noise reference (AUC = 0.618) is a separate seeded autoencoder fit within this sweep, not the primary

TABLE 6 Paired held-out comparisons: 95% class-stratified-bootstrap confidence intervals for metric differences (full model minus comparator; 2,000 paired resamples), with exact two-sided McNemar tests.

Comparison	Accuracy difference	F1 difference	FAR difference	AUC difference	McNemar *p*
FedPRISM (full)—Random Forest	[−0.020, 0.051]	[−0.020, 0.073]	[0.000, 0.054]	[−0.018, 0.010]	1.000
FedPRISM—centralized FedPRISM comparator	[−0.061, 0.040]	[−0.094, 0.049]	[−0.054, 0.000]	[−0.068, 0.000]	1.000
FedPRISM (full)—FedPRISM without Big Five	[−0.061, 0.061]	[−0.091, 0.062]	[−0.107, 0.036]	[−0.002, 0.064]	1.000

Intervals spanning zero do not demonstrate equivalence.

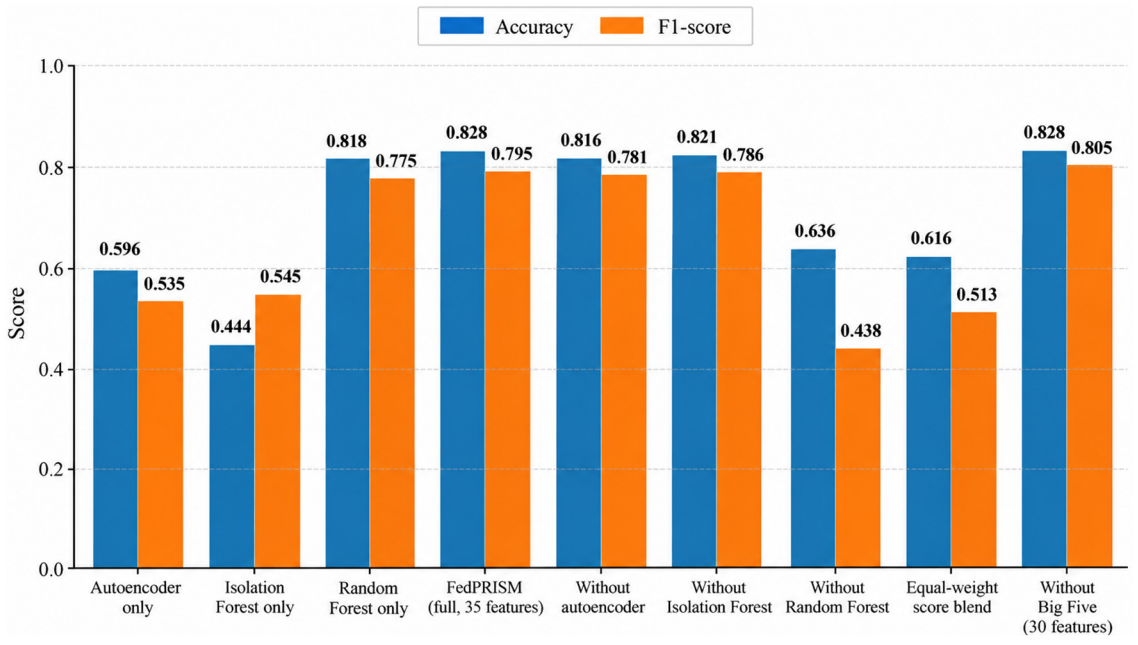

FIGURE 15
Held-out accuracy and F1 across the detector, stacking, and personality ablations. The Random Forest is the principal predictive component; the full stack is not statistically distinguishable from it on this test set.

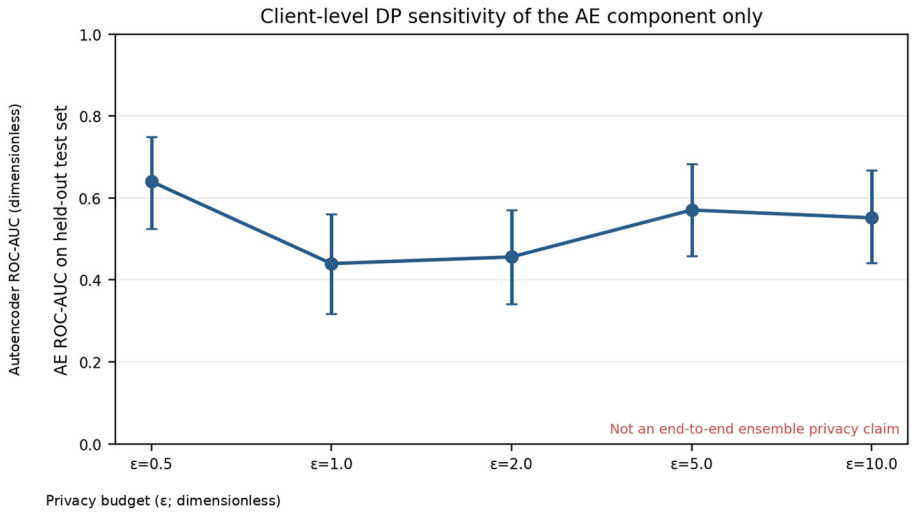

FIGURE 16
Autoencoder-only held-out ROC-AUC under the client-level DP noise settings, with stratified-bootstrap test-set intervals. This is a component sensitivity analysis, not an end-to-end privacy–utility curve for the full ensemble.

TABLE 7 Qualitative comparison of scope.

Study	Federated setting	Personalization/guidance	Privacy and evaluation scope
Aslam et al. (2024)	No	Big-Five traits; CGPA regression compared with and without traits	Centralized prediction; no DP
Takami et al. (2023)	No	Personality-tailored quiz explanations	18-day engagement evaluation; no DP
Zhong (2022)	No	Collaborative-filtered student-management recommendations	Centralized; no privacy mechanism
Afzaal et al. (2024)	No	Explainable self-regulation feedback	Course-level evaluation; no DP method
Yoneda et al. (2025)	Federated study	No recommendation module evaluated	Differential features (not DP); cross-course risk ranking
Chen and Qi (2025)	Federated DP FedAvg	No recommendation module evaluated	Programming-course prediction; DP method
FedPRISM (this study)	**Four-client simulated IID**	**Big Five in full model; no-Big-Five ablation; prompts not user-tested**	**Separate AE-only client-level DP sensitivity; no secure aggregation; retrospective held-out test**

Entries summarize the cited studies and are not a head-to-head performance ranking. Bold formatting highlights the present study (FedPRISM) to distinguish its methodological scope and contributions from those of the previously published studies.

pipeline autoencoder (AUC = 0.580). The bootstrap intervals overlap broadly, so these runs do not establish that privacy noise preserves or improves utility; they illustrate sensitivity under this small-client setup. All AUCs are for the autoencoder score alone, not the full stacked ensemble, and the intervals condition on one fitted run per setting rather than quantify variation across DP seeds.

4.6 Preliminary guidance prototype

The rule-based module can map high reconstruction error or a below-reference value on a modifiable behavior to a candidate prompt, for example discussing attendance barriers or a feasible study plan with an adviser. We do not report an individual student profile or claim that such prompts are effective interventions. The reference values are cohort associations, and the code has not been evaluated with students or educators. Accordingly, recommendations are preliminary decision-support examples requiring human review, consent-sensitive practice, and prospective validation before any use in student-facing or institutional decisions.

4.7 Comparison with existing studies

Table 7 gives a qualitative, non-head-to-head comparison with recent work. Three distinctions are relevant. First, Aslam et al. (2024) compared CGPA regression with and without Big-Five predictors, while Takami et al. (2023) evaluated personality-tailored quiz explanations. FedPRISM retains Big-Five traits in

its primary 35-feature model, adds a 30-feature no-Big-Five ablation for retrospective binary risk classification, and supplies behavior-focused candidate prompts that have not been tested as recommendations or interventions. These tasks and outcomes should not be conflated. Second, prior recommendation and feedback systems such as Zhong (2022) and Afzaal et al. (2024) differ from the present federated-learning simulation; its four clients are simulated IID shards within one institution, not independent departments or deployed cross-silo partners. We therefore do not claim that this experiment demonstrates operational data-locality across institutions. Third, Yoneda et al. (2025) use differential features rather than a formal DP mechanism, whereas Chen and Qi (2025) study a DP FedAvg method for programming-course prediction. FedPRISM's separate privacy analysis covers only the autoencoder updates and does not implement secure aggregation or protect the complete pipeline. Because datasets, targets, privacy scopes, and validation protocols differ, reported scores are not directly rankable. The contribution here is a transparent held-out comparison with FAR, detector and personality ablations, and an explicitly scoped privacy sensitivity analysis–not a claim of end-to-end privacy or universal superiority.

4.8 Limitations

This study has two main limitations. First, the data used in the present study, SAPEx-D, is limited to student observations from a single university, which might make the present results less generalizable for other universities and student groups.

Second, the ablation analysis did not reveal any incremental predictive benefit by the Big-Five personality features. However, they could differ between datasets and educational settings. Future research should examine their predictive value in conjunction with relevant family and financial characteristics as well as ensuring proper data governance, fairness, and responsible use of sensitive student information.

5 Conclusion

This study presented FedPRISM as a leakage-controlled, four-client IID federated simulation for retrospective at-risk classification and preliminary academic guidance on SAPEx-D. The 35-feature stack achieved accuracy 0.828, F1 = 0.795, FAR = 0.125, and ROC-AUC = 0.889 on a 99-student test set. Its strongest single comparator, the Random Forest, was similar; paired uncertainty did not establish a performance difference. Ablations indicated that the Random Forest supplied most of the predictive signal, while removing the five personality features did not produce a statistically clear change. Accordingly, the experiment establishes neither ensemble superiority over the Random Forest nor an incremental benefit from personality inputs.

The privacy experiment was limited to client-level DP noise on the FedAvg autoencoder parameters. It did not protect the complete ensemble, recommendation outputs, or an untrusted coordinator, and no cryptographic secure aggregation was implemented. The recommendation component provides candidate, behavior-focused prompts for human review; it has not been evaluated as an intervention. Because SAPEx-D is single-institution and cross-sectional for this experiment, the results do not establish prospective warning time, generalization to other institutions, or educational benefit.

At the evaluated operating point, the stack flagged seven of 56 on-track test records. For any future validation, this indicates a meaningful false-alert review burden: institutions would need to set thresholds in light of local advising capacity and the cost of missed cases, keep a trained human-review step, and avoid treating scores as student labels or grounds for automated action. Because warning lead time and student outcomes were not measured, live early-warning or intervention use is not justified by this experiment.

Future work should validate the pipeline on larger multi-institution cohorts with actual and heterogeneous client partitions; evaluate an end-to-end threat model with secure aggregation and privacy accounting for every released component; repeat model evaluation across independent cohorts and temporal splits; and conduct governed user studies to assess false-alert burden, fairness, recommendation acceptability, and student outcomes. Until those studies are completed, FedPRISM should be understood as a reproducible research prototype rather than a deployable early-warning or intervention system.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: the dataset is confidential and contains student-related personal, family, academic, and personality information. Access to the complete SAPEx-D dataset is restricted and can be obtained from the dataset authors upon reasonable request for research purposes. Requests to access these datasets should be directed to Muhammad Adnan Aslam, malikmadnanaslam@gmail.com.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

MAA: Conceptualization, Data curation, Methodology, Software, Writing – original draft, Writing – review & editing. MKA: Formal analysis, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. AY: Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing. ZA: Project administration, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Acknowledgments

The authors acknowledge the team responsible for the collection and curation of the SAPEx-D dataset.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Afzaal, M., Nouri, J., Zia, A., Papapetrou, P., Fors, U., Wu, Y., et al. (2021). "Automatic and intelligent recommendations to support students' self-regulation," in *Proceedings of the IEEE 21st international conference on advanced learning technologies (ICALT)*, 336–338.

Afzaal, M., Zia, A., Nouri, J., and Fors, U. (2024). Informative feedback and explainable AI-based recommendations to support students' self-regulation. *Technol. Knowl. Learn.* 29, 331–354. doi: 10.1007/s10758-023-09650-0

Albreiki, B., Zaki, N., and Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. *Educ. Sci.* 11:552. doi: 10.3390/educsci11090552

Al-Tameemi, G., Xue, J., Ali, I. H., and Ajit, S. (2024). A hybrid machine learning approach for predicting student performance using multi-class educational datasets. *Procedia Comput. Sci.* 238, 888–895. doi: 10.1016/j.procs.2024.06.108

Anupama, V., and Elayidom, M. S. (2022). "Course recommendation system: collaborative filtering, machine learning and topic modelling," in *Proceedings of the 8th international conference on advanced computing and communication systems* (Coimbatore), 1459–1462.

Aslam, M. A., Murtaza, F., Ehatisham Ul Haq, M., Yasin, A., and Ali, N. (2025). SAPEx-D: a comprehensive dataset for predictive analytics in personalized education using machine learning. *Data* 10:27. doi: 10.3390/data10030027

Aslam, M. A., Murtaza, F., Ehatisham Ul Haq, M., Yasin, A., and Azam, M. A. (2024). A human-centered approach to academic performance prediction using personality factors in educational AI. *Information* 15:777. doi: 10.3390/info15120777

Ceyhan, M., Okyay, S., Kartal, Y., and Adar, N. (2021). "The prediction of student grades using collaborative filtering in a course recommender system," in *Proceedings of the 5th international symposium on multidisciplinary studies and innovative technologies (ISMSIT)* (Ankara), 177–181.

Chen, S., and Qi, X. (2025). Entropy-adaptive differential privacy federated learning for student performance prediction and privacy protection: a case study in Python programming. *Front. Artif. Intell.* 8:1653437. doi: 10.3389/frai.2025.1653437

Dwork, C., and Roth, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* 9, 211–407. doi: 10.1561/0400000042

Embarak, O. (2021). "Towards an adaptive education through a machine learning recommendation system," in *Proceedings of the 3rd international conference on artificial intelligence in information and communication (ICAIIC)* (Jeju Island), 187–192.

Embarak, O. H., and Hawarna, S. (2024). Enhancing student success with XAI-powered RADAR: an automated AI-driven system for early detection of at-risk students. *Procedia Comput. Sci.* 231, 151–160. doi: 10.1016/j.procs.2023.12.187

Fahd, K., and Miah, S. J. (2023). Designing and evaluating a big data analytics approach for predicting students' success factors. *J. Big Data* 10:159. doi: 10.1186/s40537-023-00835-z

George, B., and Wooden, O. (2023). Managing the strategic transformation of higher education through artificial intelligence. *Adm. Sci.* 13:196. doi: 10.3390/admsci13090196

Hanaysha, J. R., Shriedeh, F. B., and In'airat, M. (2023). Impact of classroom environment, teacher competency, information and communication technology resources, and university facilities on student engagement and academic performance. *Int. J. Inf. Manag. Data Insights* 3:100188. doi: 10.1016/j.jjimei.2023.100188

Huang, Q., and Chen, J. (2024). Enhancing academic performance prediction with temporal graph networks for massive open online courses. *J. Big Data* 11:52. doi: 10.1186/s40537-024-00918-5

Kandati, D. R., and Anusha, S. (2023). Security and privacy in federated learning: a survey. *Trends Comput. Sci. Inf. Technol.* 8, 029–037. doi: 10.17352/tcsit.000066

Khan, I., Ahmad, A. R., Jabeur, N., and Mahdi, M. N. (2021). An artificial intelligence approach to monitor student performance and devise preventive measures. *Smart Learn. Environ.* 8:17. doi: 10.1186/s40561-021-00161-y

Li, H., Ge, L., and Tian, L. (2024). Survey: federated learning data security and privacy-preserving in edge-Internet of Things. *Artif. Intell. Rev.* 57:130. doi: 10.1007/s10462-024-10774-7

López-García, A., Blasco-Blasco, O., Liern-García, M., and Parada-Rico, S. E. (2023). Early detection of students' failure using machine learning techniques. *Oper. Res. Perspect.* 11:100292. doi: 10.1016/j.orp.2023.100292

Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H., and Raad, A. (2023). Reviewing federated learning aggregation algorithms: strategies, contributions, limitations and future perspectives. *Electronics* 12:2287. doi: 10.3390/electronics12102287

Pawar, M., and Dongardive, J. (2024). Improving student performance prediction through feature selection: Insights from "Offee" assessment data during the Covid-19 pandemic. *Procedia Comput. Sci.* 235, 3075–3084. doi: 10.1016/j.procs.2024.04.291

Rafi, T. H., Noor, F. A., Hussain, T., and Chae, D.-K. (2024). Fairness and privacy preserving in federated learning: a survey. *Inf. Fusion* 105:102198. doi: 10.1016/j.inffus.2023.102198

Romero, C., and Ventura, S. (2007). Educational data mining: a survey from 1995 to 2005. *Expert Syst. Appl.* 33, 135–146. doi: 10.1016/j.eswa.2006.04.005

Takami, K., Flanagan, B., Dai, Y., and Ogata, H. (2023). Personality-based tailored explainable recommendation for trustworthy smart learning system in the age

of artificial intelligence. *Smart Learn. Environ*. 10:65. doi: 10.1186/s40561-023-00282-6

Yoneda, S., Švábenský, V., Li, G., Deguchi, D., and Shimada, A. (2025). "Ranking-based at-risk student prediction using federated learning and differential features," in *Proceedings of the 18th international conference on educational data mining* (Palermo), 289–302.

Zhalgassova, Z., Shaikym, A., Sadyk, U., Kutzhan, A., Amirkumar, M., and Assangali, B. (2023). "MBTI-based recommendation system for extracurricular activities for high school students," in *Proceedings of the 17th international conference on electronics computer and computation* (Kaskelen), 1–4.

Zhao, L., Ren, J., Zhang, L., and Zhao, H. (2023). Quantitative analysis and prediction of academic performance of students using machine learning. *Sustainability* 15:12531. doi: 10.3390/su151612531

Zhong, L. (2022). "Student personalized management system based on collaborative filtering algorithm," in *Proceedings of the international conference on artificial intelligence and autonomous robot systems* (Bristol), 206–210.