ISSN: **3139-0862**

Ensemble Machine Learning for Student Performance Prediction: A Review of Methods, Evidence and the Early-Warning Gap

[1st] Rohit
MTech Scholar, Dept. Computer Science and Enginnering
Jai Narain College of Technology, Bhopal, India
rkn8742@gmail.com

[2nd] Dr. Amit Kumar Sahu
Professor, Dept. Computer Science and Enginnering
Jai Narain College of Technology, Bhopal, India Bhopal, India

[3rd] Dr. Vivek Sharma
HOD, Dept. Computer Science and Enginnering
Jai Narain College of Technology, Bhopal, India Bhopal, India

Abstract
Educational data mining has made the prediction of student academic outcomes one of its most heavily studied tasks, and ensemble learning has become the default methodological choice for it. Reported accuracies above ninety per cent are now routine, yet schools have adopted almost none of these models operationally. This review examines whether the evidence supports the confidence with which ensemble superiority is asserted, and whether the accuracies reported in the literature translate into usable early warning. Twenty-nine primary studies published between 2003 and 2025 are surveyed alongside the methodological literature on bagging, boosting, voting and stacking, and the survey is grounded in a controlled case study on the widely used UCI Student Performance corpus in which six ensemble configurations were trained under a single protocol. Three findings recur. First, the reported superiority of ensembles rests almost entirely on point estimates from single train-test splits; in the case study, no pairwise difference among six models reached significance under McNemar testing, the Friedman test returned $p = 0.145$, and a logistic regression matched the best stacking ensemble exactly on accuracy, precision, recall and F1-score. Second, model ranking is unstable at the cohort sizes typical of this field: across ten stratified partitions of the same 395 records, accuracy varied from 0.8228 to 0.9620 and the leading model changed. Third, and most consequentially, the headline accuracies depend on interim grades drawn from the same course as the target. Withdrawing them cost the strongest model 22.79 percentage points, dropping four of six configurations below the majority baseline and reducing the area under the curve to between 0.515 and 0.623.. The review argues that progress now depends less on algorithm selection than on three neglected areas: timely behavioural instrumentation to replace the interim-grade shortcut, routine statistical and multi-partition reporting, and fairness auditing of models that use demographic attributes to allocate remedial support.

Keywords—Educational data mining, student performance prediction, ensemble learning, stacking, gradient boosting, random forest, explainable artificial intelligence, SHAP, early-warning systems, learning analytics.

1 INTRODUCTION

1.1 Overview

Schools have collected data about their students for as long as formal schooling has existed. What changed over the past two decades is that these records moved into digital systems, and that their volume now exceeds what a class teacher can process by inspection. Educational data mining (EDM) emerged as the response to that surplus: the application of computational discovery methods to data produced in educational settings, with the practical aim of understanding learners and the environments in which they learn [1].

Successive surveys record a shift from descriptive analysis, which asks what happened, towards predictive modelling, which asks what is likely to happen and to whom [2], [3]. Prediction of academic performance is now among the most heavily studied tasks in the field. The outcome variable is administratively unambiguous, the data are comparatively cheap to obtain, and the intended intervention is obvious: a student flagged as likely to fail can be offered remedial support while support can still change the result.

Ensemble learning supplies most of the modelling apparatus. Rather than committing to one hypothesis, an ensemble combines several learners whose collective decision is generally more stable than that of any member [4]. Bagging reduces variance by training members on resampled data [5], [6], boosting reduces bias by fitting members sequentially to the errors of their predecessors [7]-[9], and stacking learns from data how much to trust each member [10]. Efficient implementations such as XGBoost and LightGBM made these ideas routine on tabular data [11], [12]. Fig. 1 summarises the resulting design space as it appears in the educational literature.

ISSN: 3139-0862

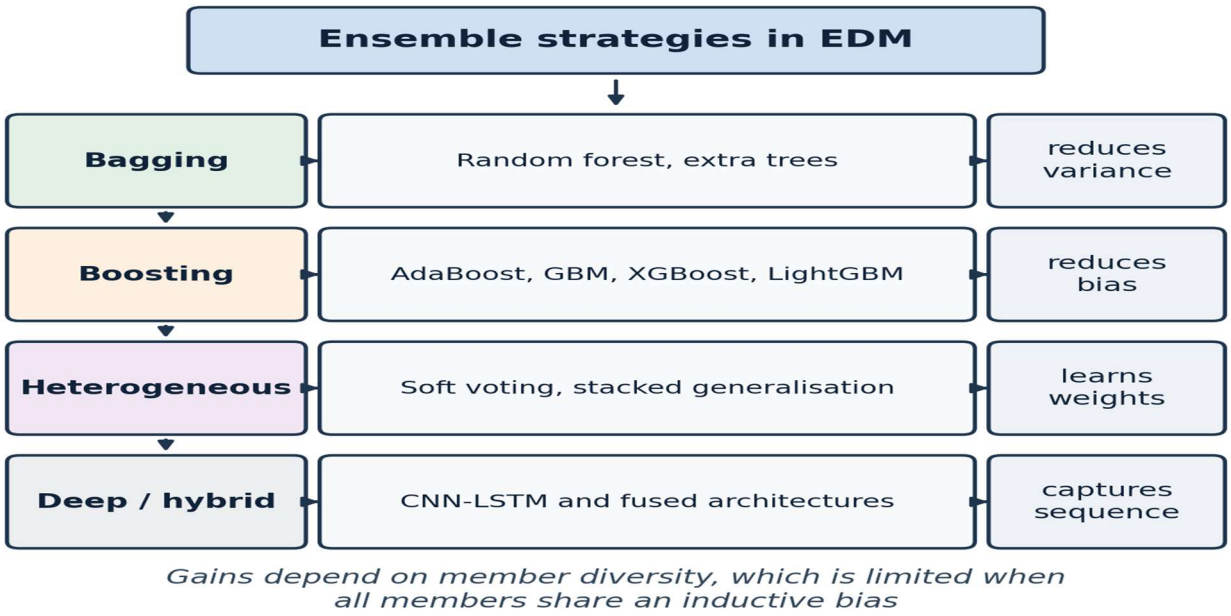


Fig. 1. Taxonomy of ensemble strategies encountered in the educational data mining literature, with the error component each family principally targets.

1.2 Importance and Motivation

A striking asymmetry motivates this review. The literature reports accuracies that would be considered excellent in most applied domains, frequently above ninety per cent, yet operational adoption in schools remains rare. Three observations suggest that the gap is not merely one of engineering effort.

The first concerns what the models are actually using. A large share of work on secondary-school cohorts supplies interim grades from the same course as predictors of the final grade. Because the target is derived from a later assessment in that same course, this is close to predicting a quantity from its own recent history. The authors who released the most widely used benchmark noted the dependency explicitly and reported results with and without those attributes [13]; much subsequent work reports only the favourable configuration.

The second concerns statistical practice. Comparative studies commonly rank six or eight models by accuracy on a single partition and declare the leading entry superior. On a test partition of eighty students, one reclassified student moves accuracy by 1.25 percentage points, which exceeds most of the differences being reported. Procedures for deciding whether such orderings survive resampling have been available for decades [14], [15] and are applied in a minority of the studies surveyed here.

The third concerns deployment. Secondary schools rarely employ data scientists. A model that cannot be reproduced from public data with open tooling, and that cannot explain why a particular student was flagged, is unlikely to be adopted whatever its accuracy. This has driven interest in attribution methods that decompose an individual prediction into per-attribute contributions [16], [17], and a parallel argument that interpretable models should be preferred outright for high-stakes decisions [18].

1.3 Objectives of the Review

This review pursues five objectives. To organise the methodological literature on ensemble learning as it is applied to student performance prediction, and to state precisely which error component each family addresses.

a) To survey and classify the primary empirical studies, recording dataset, method family, evaluation protocol and reported outcome in a form that permits comparison.
b) To assess the quality of evidence for the claim that ensembles outperform single learners, distinguishing what has been demonstrated from what has been asserted.
c) To quantify, through a controlled case study executed under a single protocol, the dependence of reported accuracy on interim assessment data.
d) To identify the technical and methodological gaps that stand between the current literature and a deployable early-warning system, and to propose directions that follow from them.
e) The contribution intended is not a further leaderboard entry. It is an assessment of how much the existing leaderboard actually establishes, supported by an experiment designed to test the two assumptions on which it rests.

2 REVIEW METHODOLOGY

Sources were located through the digital libraries of IEEE, Elsevier, Springer, ACM and MDPI, using combinations of the terms student performance

ISSN: **3139-0862**

prediction, educational data mining, learning analytics, ensemble learning, at-risk student and early warning. The search was supplemented by backward and forward citation tracking from the field surveys [1]-[3] and by the reference lists of the retained articles.

A record was retained as a primary study if it reported a quantitative evaluation of at least one supervised learner on data collected in an educational setting, stated its evaluation metric, and described its dataset sufficiently for the result to be interpreted. Records describing only a system architecture, position papers without evaluation, and studies whose target was not an academic outcome were excluded. Priority was given to work published from 2016 onwards, with earlier material retained where it introduced a method or a framing still in use.

Twenty-nine primary studies were retained and classified by dominant method family. These are supported by three field surveys and by thirty-five methodological, statistical and tooling references. Fig. 2 records the selection process and Fig. 3 the composition of the retained corpus. The review is organised thematically rather than chronologically, because the field's development is better described by shifts in method family than by year.

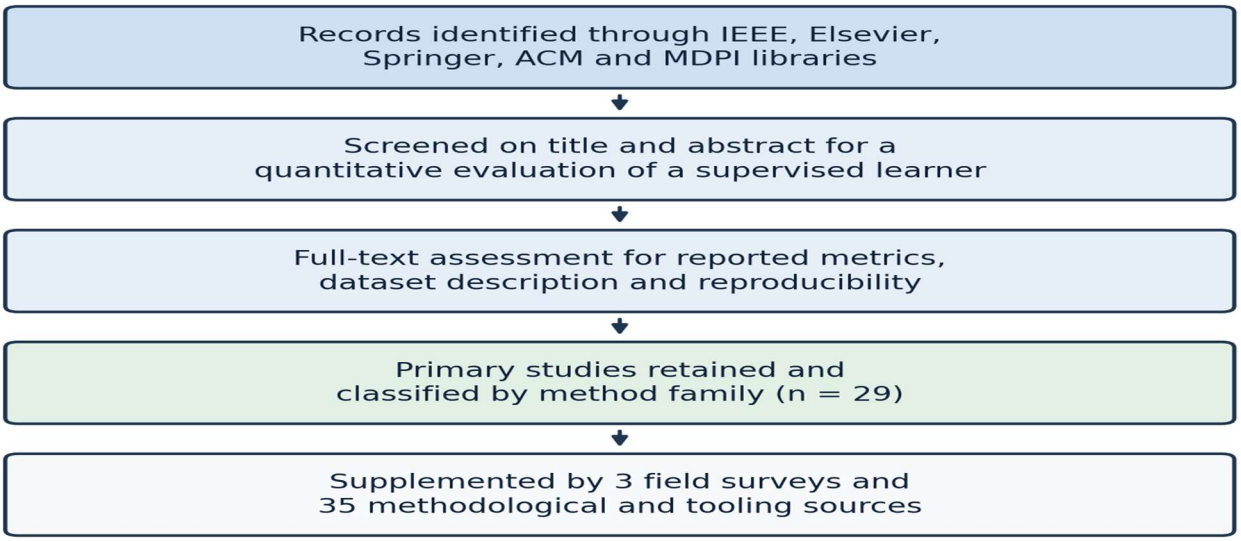


Fig. 2. Identification and screening of sources. The retained corpus comprises 29 primary studies classified by method family.

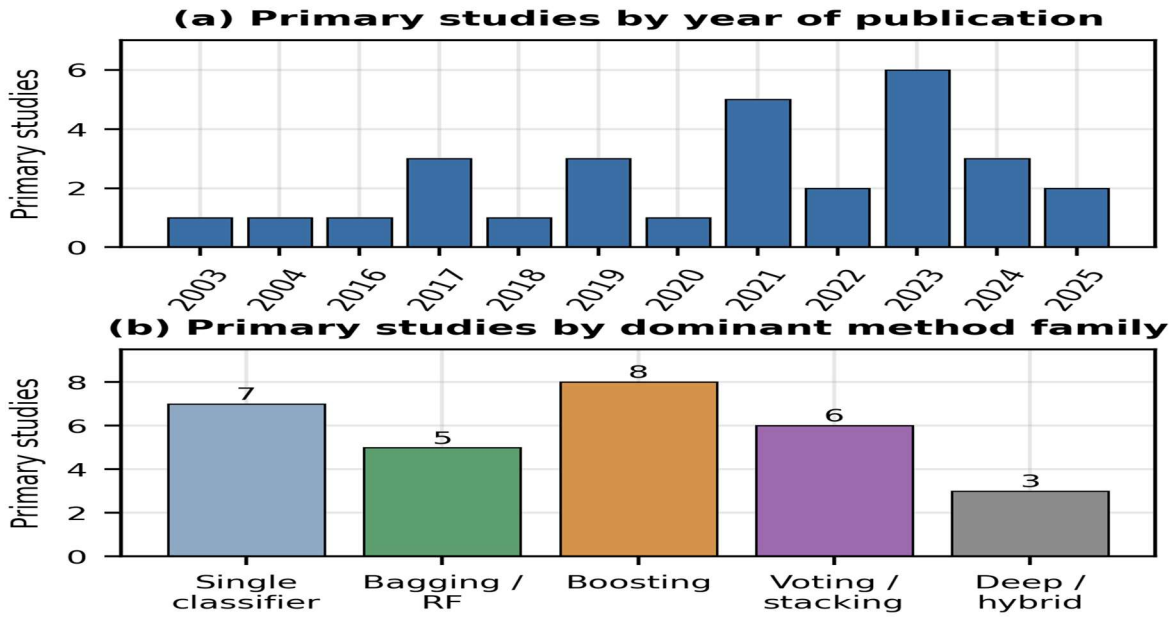


Fig. 3. Composition of the reviewed corpus: (a) primary studies by year, showing the concentration of activity after 2016; (b) primary studies by dominant method family.

ISSN: **3139-0862**

3 THEORETICAL FOUNDATIONS

Let a cohort be represented by $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \in \mathbb{R}^p$ describes one student and $y_i \in \{0, 1\}$ records the outcome, with 1 denoting a pass. The pairs are assumed drawn independently from an unknown joint distribution P(x, y), and the task is to select $h : \mathbb{R}^p \to \{0, 1\}$ minimising the expected loss given in (1).

$$R(h) = \mathbb{E}_{(x,y)\sim P}\,[\, L(h(x), y)\,] \qquad (1)$$

Since P is unknown the risk is estimated on a finite sample. The expected error of a learner decomposes into squared bias, variance and irreducible noise [19], and the two dominant ensemble families are best understood as attacking different terms of that decomposition.

A. Bagging and the Random Forest

Bootstrap aggregation draws B samples of size n with replacement, fits a model on each, and votes [5]. If members were independent with variance σ^2, the variance of their average would be σ^2/B. Members trained on overlapping samples are not independent, and with pairwise correlation ρ the variance of the average is (2).

$$\mathrm{Var}(\hat{h}_{avg}) = \rho\sigma^2 + (1 - \rho)\,\sigma^2 / B \qquad (2)$$

The first term does not vanish as B grows, so the achievable reduction is bounded by the correlation between members. The random forest attacks that term directly by considering only a random subset of $m \le p$ attributes at each split, decorrelating the trees at the cost of a small increase in individual bias [6]. The idea has a precursor in the random subspace method [20] and a more aggressive variant in extremely randomised trees [21]. Accumulated impurity decrease yields the familiar importance measure (3), where v(t) is the attribute split at node t.

$$\mathrm{Imp}(j) = (1/B)\,\Sigma_{b=1}^{B}\,\Sigma_{t:\,v(t)=j}\,(n_t / n)\,\Delta I(t) \qquad (3)$$

3.1 Boosting

Boosting builds its committee sequentially, each member concentrating on instances its predecessors handled badly. Schapire established that a weak learner can be converted into a strong one [7], and AdaBoost gave the idea a practical form by reweighting instances after each round [8]. Friedman recast the procedure as gradient descent in function space [9]: the model is additive as in (4), and each new member is fitted to the negative gradient of the loss at the current model, (5).

$$F_M(x) = F_0(x) + \Sigma_{m=1}^{M}\,\nu\,\gamma_m\,h_m(x) \qquad (4)$$

$$r_{i,m} = -\,[\,\partial L(y_i, F(x_i)) / \partial F(x_i)\,]_{F = F_{m-1}} \qquad (5)$$

XGBoost augments this with explicit regularisation and a second-order expansion of the loss, giving the split gain criterion (6), in which G and H are summed gradients and Hessians and γ acts as pre-pruning [11].

$$\mathrm{Gain} = \tfrac{1}{2}\,[\,G_L^2/(H_L+\lambda) + G_R^2/(H_R+\lambda) - (G_L+G_R)^2/(H_L+H_R+\lambda)\,] - \gamma \qquad (6)$$

LightGBM targets the cost of split finding through gradient-based one-side sampling and exclusive feature bundling, and grows trees leaf-wise rather than level-wise, which lowers training loss faster but overfits more readily on small cohorts [12]. CatBoost addresses categorical attributes through ordered target statistics and has been applied to secondary-school data with parameter search [22].

3.2 Heterogeneous Combination

Soft voting averages the predicted probabilities of M members as in (7) and thresholds the average. It is preferred to hard voting when members are calibrated, but it weights every member equally regardless of competence.

$$\hat{p}_{soft}(x) = (1/M)\,\Sigma_{m=1}^{M}\,\hat{p}_m(x) \qquad (7)$$

Stacking replaces the fixed rule with a learned one [10]. Level-zero members produce predictions that become the inputs of a level-one meta-learner, as in (8), where σ is the logistic function.

$$\hat{p}_{stack}(x) = \sigma(\,\beta_0 + \Sigma_{m=1}^{M}\,\beta_m\,\hat{p}_m(x)\,) \qquad (8)$$

The essential implementation detail is that the meta-features must be generated out of fold. If level-zero members were fitted on the whole training set and their in-sample predictions passed upward, the meta-learner would observe artificially confident inputs and learn to trust an over-fitted member. Several published pipelines omit this safeguard, and where it is omitted the reported gain from stacking is not interpretable.

3.3 Why Ensembles Help, and When They Do Not

Ensemble gains depend on diversity: if members make identical errors, averaging changes nothing. Kuncheva and Whitaker examined ten diversity measures and found the relationship between diversity and ensemble accuracy weaker and less monotone than intuition suggests [23]. Zhou reaches a similar conclusion, that diversity is necessary but not sufficient and must not be bought by sacrificing member accuracy [24]. Sagi and Rokach note that heterogeneous ensembles tend to be more diverse than homogeneous ones [25].

This has a direct bearing on the educational case. Where one or two attributes dominate the signal, as they do in most of the cohorts reviewed in Section IV, every tree-based member is fitting essentially the same function. Diversity is limited from the outset, and modest ensemble gains are the expected outcome rather than a disappointing one. Table I summarises the properties that distinguish the families.

ISSN: **3139-0862**

Table I. Characteristics of the ensemble families applied to student performance prediction.

Property	Bagging	Boosting	Stacking
Member training	Parallel	Sequential	Parallel, then meta-level
Error term reduced	Variance	Bias	Both, indirectly
Member depth	Deep	Shallow	Depends on members
Combination rule	Fixed vote	Weighted additive	Learned
Noise sensitivity	Low	Higher	Inherited
Parallelisable	Fully	Limited	Level zero only
Principal risk	Correlated members	Late-round overfitting	Leakage if not out of fold

4 LITERATURE REVIEW

4.1 Single-Classifier Foundations

The earliest systematic comparisons predate the ensemble era. Minaei-Bidgoli and colleagues applied decision trees, neural networks and nearest-neighbour classifiers to data from a web-based physics course and reported that combining classifiers improved on any single method, an early signal of what later became the dominant strategy [26]. Kotsiantis and colleagues addressed dropout in distance education and found naive Bayes competitive with more elaborate learners when the feature set was small [27]. That result is worth recalling, because it is the first published instance of a pattern this review encounters repeatedly: model complexity is not automatically rewarded in this domain.

Later single-classifier work grew more careful about timing. Marbouti and colleagues built models for a first-year engineering course that could be applied at successive points in the term and showed that accuracy at week four differs materially from accuracy at week twelve [28]. Xu and colleagues developed a progressive framework that updates its estimate of degree performance as new course results arrive, treating prediction as sequential rather than one-shot [29]. Asif and colleagues combined prediction with clustering across four undergraduate cohorts, arguing that a purely predictive output gives an adviser too little to act on [30]. Zabriskie and colleagues applied logistic regression and random forests to introductory physics enrolment data and found prior academic record dominating demographic attributes [31], a finding echoed in nearly every study surveyed here.

Yağcı compared six classifiers on 1854 undergraduates using midterm marks and departmental data alone, obtaining accuracies between seventy and seventy-five per cent [32]. This figure anchors the later discussion: it represents what is achievable when only limited academic history is supplied, and it sits far below the accuracies reported where interim grades from the target course are available.

4.2 Homogeneous Ensembles

Dietterich set out the theoretical case for ensembles in three parts: a statistical argument, when the data are insufficient to identify the best single hypothesis; a computational argument, when the learner finds only local optima; and a representational argument, when the true function lies outside the hypothesis space of any member [4]. All three apply to educational data, and the first applies with particular force to cohorts of a few hundred records.

Injadat and colleagues developed a multi-split optimised bagging framework for multi-class educational data mining, evaluating it on two undergraduate datasets at different points of course delivery and reporting consistent gains over single learners [33]. Moubayed and colleagues surveyed the same territory from an engineering perspective and identified the constraint that shapes the field: educational datasets are small, heterogeneous in attribute type, and collected under conditions that make cross-institutional pooling difficult [34].

Boosting has attracted the most attention. Asselman and colleagues integrated XGBoost into a performance factors analysis framework and reported improvements over the standard formulation across three datasets [35]. Duan and colleagues applied XGBoost directly and emphasised its tolerance of mixed attribute types [36]. Fan and colleagues optimised a categorical boosting model with a double particle swarm procedure specifically for secondary-school performance, the closest published analogue

ISSN: **3139-0862**

to the case study reported in Section VI [22]. Chen and Zhai carried out a broad comparison across learners and datasets that includes the same benchmark corpus [37], and Hoyos and Caicedo-Castro examined how parameter tuning alters predictive quality on it [38].

4.3 Heterogeneous Ensembles

Heterogeneous combination is the more recent and faster-growing strand. Karalar and colleagues assembled an ensemble to identify students at risk of failure during pandemic-period distance learning, reporting specificity above ninety per cent on a cohort of more than two thousand [39]. Keser and Aghalarova proposed a hybrid algorithm that selects its constituent learners adaptively rather than fixing the committee in advance [40]. Butt and colleagues built a multi-model ensemble for higher education and coupled it with an analysis of assessment environments [41]. Priyambada and colleagues constructed a two-layer ensemble fusing learning behaviour with domain knowledge, an architecture close in spirit to stacking [42]. Cetin and Cetin applied stacking to large-scale international assessment data and found it superior to its base learners across several country subsamples [43], one of the few studies in which sample size is large enough for such a comparison to carry weight.

Adnan and colleagues examined at-risk prediction at successive fractions of course length using the Open University Learning Analytics dataset and found, as Marbouti had, that achievable accuracy rises steadily with the proportion of the course completed [44], [45]. The consistency of this finding across institutional contexts is one of the more robust results in the field, and Section V returns to its implications.

4.4 Comparative Summary

Table II summarises the studies most directly comparable to the case study of Section VI. Two columns deserve comment. The evaluation column records whether any inferential procedure was applied to the reported differences; in the majority of cases the answer is that none was. The interim-data column records whether assessment results from the target course itself were supplied as predictors, which Section VI shows to be the single largest determinant of reported accuracy.

Read as a whole, the table shows a field converging on a narrow band of reported performance, roughly 0.85 to 0.93 accuracy where interim grades are available and roughly 0.65 to 0.78 where they are not, largely irrespective of which ensemble is used. That convergence is itself informative. It suggests that the ceiling is set by the informativeness of the recorded attributes rather than by the sophistication of the learner.

Table II. Comparative summary of representative primary studies on student performance prediction. "Inference" records whether any statistical test was applied to the reported differences; "Interim" records whether assessment results from the target course were supplied as predictors.

Study	Year	Cohort and size	Dominant method	Reported outcome	Inference	Interim
Minaei-Bidgoli et al. [26]	2003	Web-based physics course	DT, NN, kNN, combination	Combination beat single learners	No	Yes
Kotsiantis et al. [27]	2004	Distance learning, 350	Naive Bayes, DT, NN	NB competitive on small feature sets	No	Partial
Cortez and Silva [13]	2008	Portuguese secondary, 395 and 649	DT, RF, NN, SVM	~0.90 with grades; ~0.67 without	No	Both reported
Marbouti et al. [28]	2016	First-year engineering	LR, DT, NB, SVM	Accuracy rises through the term	No	Partial
Asif et al. [30]	2017	Undergraduate, four cohorts	DT, NB, NN with clustering	Prior marks dominate	No	Yes
Xu et al. [29]	2017	Degree programmes	Progressive ensemble	Sequential updating effective	No	Partial
Moubayed et al. [34]	2018	Survey of e-learning analytics	Survey	Small heterogeneous data is the constraint	n/a	n/a
Zabriskie et al. [31]	2019	Introductory physics	LR, random forest	Prior record beats demographics	Partial	Yes
Injadat et al. [33]	2020	Two undergraduate sets	Optimised bagging	Gains over single learners	No	Yes
Adnan et al. [44]	2021	OULAD, large	RF, SVM, LR and others	Accuracy grows with course fraction	No	Partial
Karalar et al. [39]	2021	Distance learning, 2000+	Voting ensemble	Specificity above 0.90	No	Activity data
Keser and Aghalarova	2022	Multiple educational sets	Adaptive hybrid ensemble	Improves on fixed committees	Partial	Yes

ISSN: **3139-0862**

[40]						
Yagci [32]	2022	Undergraduate, 1854	RF, SVM, LR, NB, kNN	0.70 to 0.75 accuracy	No	Midterms only
Asselman et al. [35]	2023	Three EDM datasets	XGBoost with PFA	Improvement over standard PFA	No	Yes
Fan et al. [22]	2023	Secondary school	PSO-tuned CatBoost	~0.93 accuracy	No	Yes
Butt et al. [41]	2023	Higher education	Multi-model ensemble	Ensemble best overall	No	Yes
Priyambada et al. [42]	2023	LMS behaviour and domain data	Two-layer ensemble	Fusion beats either source	No	Behavioural
Hoyos and Caicedo-Castro [38]	2024	Same UCI corpus	Tuned classical models	0.85 to 0.92 accuracy	No	Yes
Talebi et al. [46]	2024	MOOC, large sequential	CNN-LSTM ensemble	Deep fusion effective at scale	No	Behavioural
Cetin and Cetin [43]	2025	PISA 2022, very large	Stacking	Stacking above base learners	Partial	n/a
This review (case study)	2026	Same UCI corpus, 395	RF, GBM, XGB, LGBM, voting, stacking	0.886 (S1), 0.785 (S2), 0.696 (S3)	Yes	All three

5 CURRENT TRENDS AND DEVELOPMENTS

5.1 From Static Snapshots to Temporal Models

The clearest trend is the movement away from single-snapshot prediction. Where early work fitted one model to an end-of-term feature vector, recent work models the trajectory. Xu and colleagues update predictions as course results accumulate [29]; Adnan and colleagues report accuracy as a function of elapsed course fraction [44]; knowledge-tracing approaches model the evolving state of a learner across exercises rather than summarising it [47]. The practical consequence is that a study reporting a single accuracy figure is now understood to be reporting one point on a curve, and the position of that point matters as much as its height.

5.2 Deep and Hybrid Architectures

Deep ensembles have appeared where data volume permits. Talebi and colleagues combined convolutional and recurrent architectures for dropout prediction in massive open online courses, a setting supplying tens of thousands of sequential records [46]. Malik and colleagues reported a modified ensemble for forecasting adaptability in online entrepreneurship education [48]. These architectures are not transferable to the cohort sizes typical of school-level studies, and the review found no convincing demonstration of a deep model outperforming a tuned tree ensemble on a few hundred tabular records. Reporting that absence is more useful than repeating the general enthusiasm for deep learning.

5.3 Explainability as a Requirement

The second clear trend is the treatment of explanation as a requirement rather than an optional extra. SHAP assigns each attribute a contribution derived from Shapley values, satisfying local accuracy, missingness and consistency [16]; the tree-specific algorithm computes exact values in polynomial time, making it directly applicable to the ensembles reviewed here [17]. Molnar provides a systematic treatment of these and competing methods [49]. Wang and Luo built interpretable models of academic achievement and used attribution to identify actionable factors rather than merely ranking them [50].

Rudin's counter-argument deserves weight in this domain specifically: for high-stakes decisions, an interpretable model should be preferred to a post-hoc explanation of an opaque one, because an approximately faithful explanation can mislead precisely when it matters [18]. Section VI provides an empirical datum bearing on that argument, in that a logistic regression matched the best ensemble on every thresholded metric while being interpretable by construction.

5.4 Evaluation Practice and Its Slow Improvement

Evaluation practice is improving, but unevenly. Accuracy remains the headline metric even under class imbalance, although Chicco and Jurman show that the Matthews correlation coefficient gives a more informative summary of a binary confusion matrix because it uses all four cells [51], and Saito and Rehmsmeier show precision-recall curves to be more diagnostic than receiver operating characteristic curves when the positive class is rare [52]. Class imbalance is usually addressed by resampling, most often SMOTE [53], a technique whose application to training partitions of a few hundred records deserves more scepticism than it typically receives.

Statistical comparison remains the weakest link. Dietterich analysed five tests for comparing supervised learners and recommended McNemar's

ISSN: **3139-0862**

test where a single test set is available [14], [54]; Demšar extended the analysis to multiple datasets and recommended non-parametric procedures such as the Friedman test with post-hoc analysis [15], [55]. Of the twenty-nine primary studies reviewed, a minority report any inferential procedure at all. Fig. 4 illustrates why this matters by placing reported accuracies against the availability of interim assessment data, the factor that turns out to dominate all others.

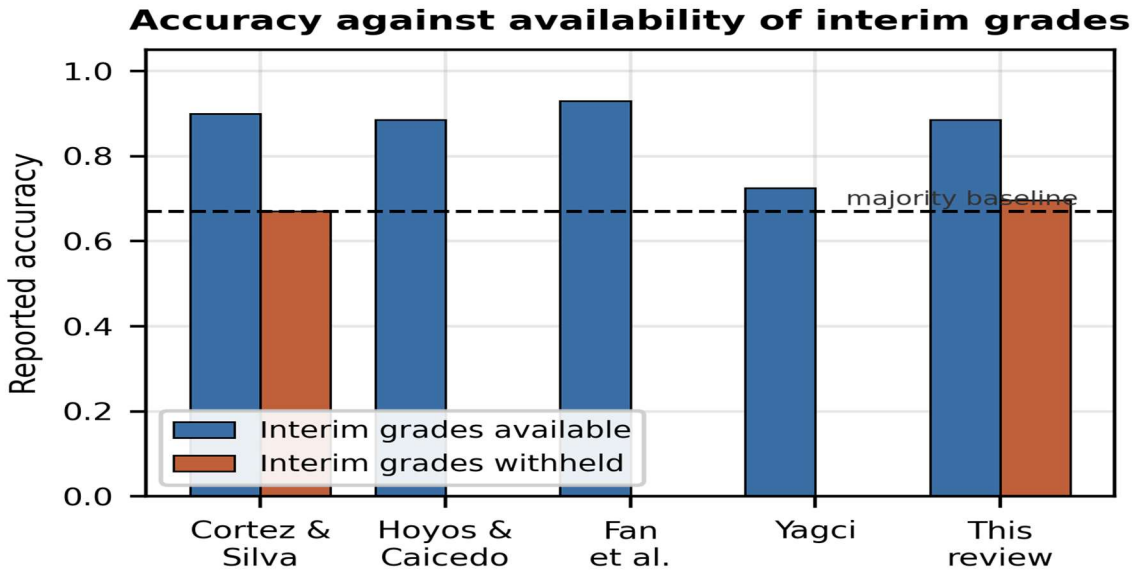


Fig. 4. Reported accuracy against the availability of interim grades. Values are as stated by the respective authors; where a range was reported, its midpoint is shown. The dashed line is the majority-class baseline of the benchmark cohort.

6 EMPIRICAL CASE STUDY

A review of comparative claims is more useful when it can test them. This section reports a controlled experiment executed for the purpose, in which six ensemble configurations were trained on a benchmark widely used in the reviewed literature, under a single protocol, with the statistical procedures that most of that literature omits.

6.1 Data and Protocol

The mathematics subset of the Student Performance corpus was used: 395 records described by 33 attributes covering demographic background, family circumstances, study habits, social behaviour and three periodic grades, collected from two Portuguese secondary schools [13], [56]. No values were missing and no duplicate records were present. The final grade was dichotomised at the official pass mark of ten on the twenty-point scale, giving 265 passing and 130 failing students, a ratio of 2.04 to 1.

Seventeen nominal attributes were label encoded and seven domain-inspired constructs were appended, among them a study-efficiency ratio, a combined alcohol-consumption index, a parental-education average and a banded absence variable, expanding the representation from 32 to 39 predictors. A stratified eighty-twenty split produced 316 training and 79 test records. Four base learners, random forest, gradient boosting, XGBoost and LightGBM, and two heterogeneous ensembles, soft voting and stacking with a logistic-regression meta-learner, were fitted with hyperparameters fixed a priori rather than tuned, so that no selection optimism could contaminate the comparison. Assessment used five-fold stratified cross-validation within the training partition and a single evaluation on the held-out partition. The implementation used scikit-learn, XGBoost, LightGBM and SHAP [57], [11], [12], [17].

6.2 Comparative Performance

Table III reports the outcome and Fig. 5 displays it. The stacking ensemble led on four of five test metrics, reaching 0.8861 accuracy, 0.9109 F1-score and 0.9608 area under the curve. Its advantage over the next configuration is 1.27 percentage points, which corresponds to exactly one additional correctly classified student.

ISSN: **3139-0862**

TABLE III. CASE STUDY: SIX CONFIGURATIONS ON THE HELD-OUT PARTITION OF 79 STUDENTS. CV COLUMNS ARE COMPUTED WITHIN THE TRAINING PARTITION.

Model	CV mean	CV std	Acc.	Prec.	Rec.	F1	AUC
Stacking	0.9368	0.0220	0.8861	0.9583	0.8679	0.9109	0.9608
Random forest	0.9177	0.0274	0.8734	0.9574	0.8491	0.9000	0.9601
Soft voting	0.9336	0.0251	0.8734	0.9388	0.8679	0.9020	0.9550
XGBoost	0.9305	0.0234	0.8734	0.9388	0.8679	0.9020	0.9419
Gradient boosting	0.9431	0.0161	0.8608	0.9200	0.8679	0.8932	0.9267
LightGBM	0.9210	0.0345	0.8608	0.9565	0.8302	0.8889	0.9514
Logistic regression	0.8988	0.0326	0.8861	0.9583	0.8679	0.9109	0.9354

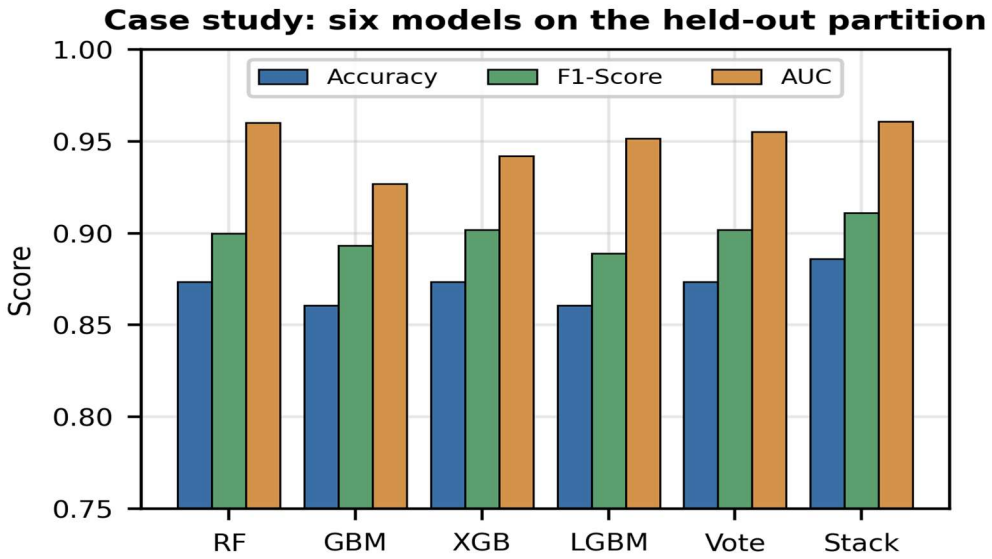


Fig. 5. Comparative performance of the six ensemble configurations on the held-out partition. Differences among them fall within one to two reclassified students.

Two observations complicate the ranking. First, the ordering by cross-validation does not match the ordering by test accuracy: gradient boosting has the highest cross-validated mean at 0.9431 and the lowest fold variance, yet ties for last on the test partition. Second, and more awkward for the ensemble thesis, a logistic regression trained on the same standardised features matched the stacking ensemble exactly on accuracy, precision, recall and F1-score, differing only in the area under the curve. Where a decision boundary is close to linear, as it is when two attributes correlate with the target at 0.725 and 0.665, eighteen seconds of stacking buys ranking quality rather than better decisions.

6.3 Statistical Significance

Exact binomial McNemar tests were applied to all fifteen model pairs on the paired test predictions. The largest discordant count anywhere in the matrix was three, and the smallest p-value obtained was 0.500. Paired t-tests on the five cross-validation folds produced a smallest p-value of 0.100, and a Friedman test across all six configurations returned a statistic of 8.209 with $p = 0.145$. On this cohort, with this protocol, no model is statistically distinguishable from any other.

This is an absence of evidence rather than evidence of equivalence: with 79 test instances and models agreeing on at least 76 of them, McNemar's test has almost no power to detect differences of the magnitude observed. That is precisely the point. The same limitation applies to the many published comparisons that report a ranking on a partition of comparable size and describe the leading entry as superior.

6.4 Stability of the Ranking

The protocol was repeated over ten independent stratified partitions of the same records. Fig. 7 shows the result. Accuracy ranged from 0.8228 to 0.9620 depending only on which students fell into the test partition, and the mean accuracies of all six configurations lay within 0.88 percentage points of one another while individual standard deviations

ISSN: **3139-0862**

were three to four times larger. Averaged over the ten partitions, XGBoost rather than stacking had the highest mean at 0.8949, reversing the single-partition ranking.

The practical reading is that model choice within this family is a second-order consideration compared with partition luck, and that any single-split result on a cohort of a few hundred records carries an uncertainty of roughly three percentage points. Differences smaller than that convey no information about the algorithms.

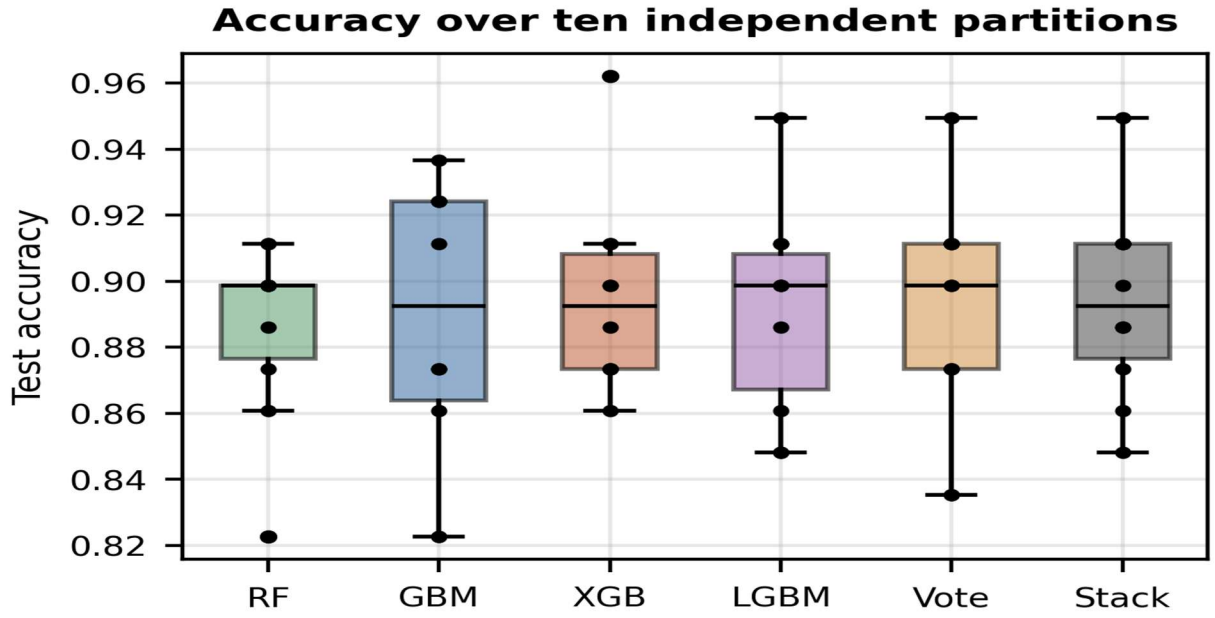


Fig. 7. Test accuracy of each configuration over ten independent stratified partitions. Dispersion across partitions exceeds the spread across models by a factor of three to four.

6.5 The Early-Warning Gap

The result with the largest practical consequence concerns the interim grades. Three feature scenarios were defined: S1 retains both interim grades, S2 withdraws the second-period grade, and S3 withdraws both together with the derived study-efficiency ratio, so that no residual grade information leaks through. All six configurations were refitted from scratch in each scenario. Table IV and Fig. 6 report the outcome.

Table IV. Test accuracy and area under the curve under three feature scenarios. S3 corresponds to a prediction made before any assessment result exists.

Model	S1 acc	S1 AUC	S2 acc	S2 AUC	S3 acc	S3 AUC
Random forest	0.8734	0.9601	0.7722	0.8810	0.6962	0.6226
Gradient boosting	0.8608	0.9267	0.7468	0.8520	0.5949	0.5152
XGBoost	0.8734	0.9419	0.7722	0.8766	0.6456	0.5610
LightGBM	0.8608	0.9514	0.7468	0.8824	0.6709	0.5537
Soft voting	0.8734	0.9550	0.7342	0.8657	0.6456	0.5522
Stacking	0.8861	0.9608	0.7848	0.8759	0.6582	0.5965

ISSN: **3139-0862**

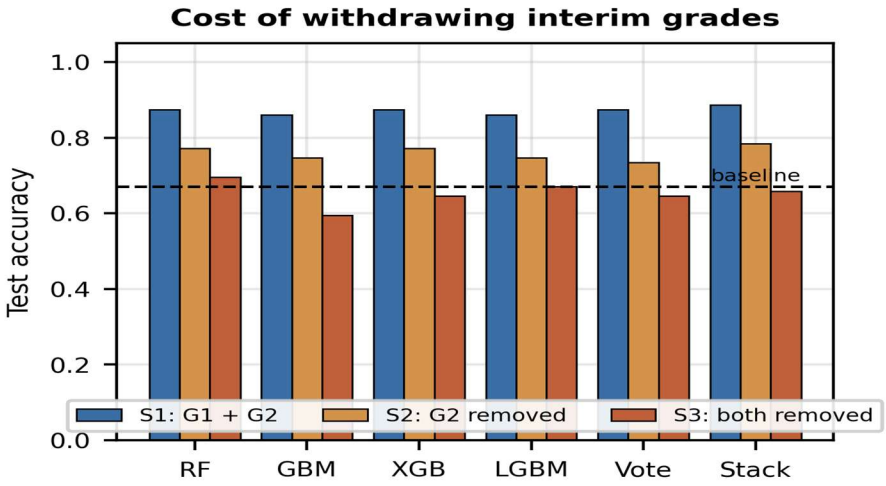


Fig. 6. Effect of withdrawing interim grades. Under S3 four of six configurations fall below the majority-class baseline shown by the dashed line.

Withdrawing the second-period grade alone cost the stacking ensemble 10.13 percentage points. Withdrawing both cost 22.79 points, leaving 0.6582. The accuracy figures understate the collapse: the majority baseline is 0.6709, so four of the six configurations perform worse than predicting pass for every student, and the area under the curve falls to between 0.5152 and 0.6226 where 0.5 is chance.

Stripped of assessment results from the target course, demographic background, family circumstances, study habits, social behaviour and past failure history together do not support useful ranking of students by risk in this cohort. The accuracies celebrated in much of the literature reviewed in Section IV rest on the observation that a student's grade in the second period predicts their grade in the third, which is not a pedagogically interesting discovery. Cortez and Silva made this point when the data were released [13]; the present experiment confirms it quantitatively for six modern ensembles sixteen years later.

One qualification tempers the pessimism. Scenario S2, corresponding to a prediction made after the first assessment period, retains accuracies of 0.73 to 0.78 and areas under the curve of 0.85 to 0.88. That is a genuinely useful operating regime, and it still leaves two thirds of the academic year in which to intervene. The correct conclusion is not that prediction is hopeless but that it becomes viable only once some assessment evidence exists.

6.6 Attribution

Impurity-based importance and SHAP attribution agreed on the three leading predictors, the two interim grades and the derived study-efficiency ratio, which together carry 66.0 per cent of impurity importance. Below that they diverge in an explicable way, shown in Fig. 8: SHAP promotes attributes that affect many students slightly, such as recorded absences and weekend alcohol consumption, whereas impurity rewards attributes that affect few students decisively, such as past class failures, which is zero for 79 per cent of the cohort. The four tree learners also disagreed sharply about concentration, gradient boosting placing 76.7 per cent of its importance on a single attribute against LightGBM's 15.5 per cent. Reporting importance from one learner, as most reviewed studies do, therefore understates how model-dependent such rankings are.

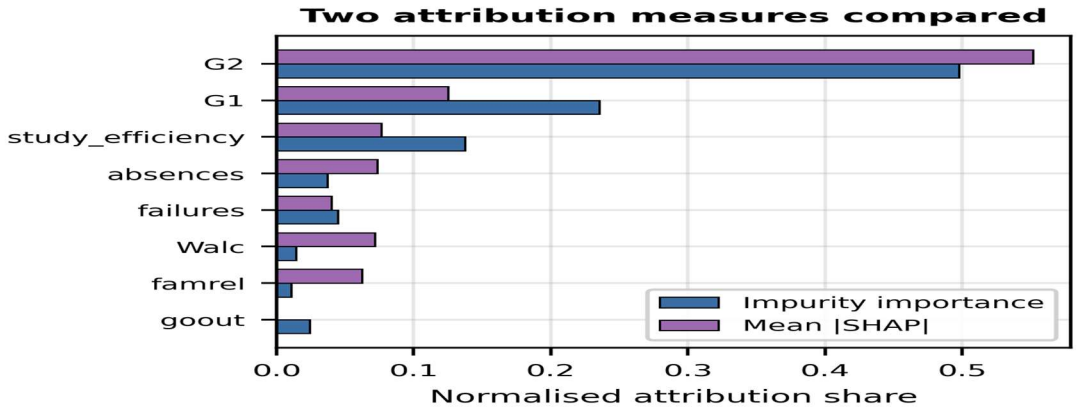


Fig. 8. Impurity-based importance compared with mean absolute SHAP attribution for the eight leading predictors. The two measures agree at the top and diverge below it.

ISSN: **3139-0862**

7 CHALLENGES AND LIMITATIONS

7.1 Technical Issues

Small and heterogeneous cohorts. Educational datasets are typically a few hundred to a few thousand records with mixed attribute types [34]. This bounds what any learner can extract, defeats deep architectures, and renders statistical comparison underpowered, as Section VI-C demonstrated concretely.

Target leakage through assessment history. The dominant technical problem is not overfitting in the usual sense but the use of predictors that are near-restatements of the target. This inflates every reported metric simultaneously, so internal validation does not detect it. Only an explicit ablation reveals the dependency, and few studies perform one.

Probability quality. Boosted models in the case study produced log losses up to 0.78 against 0.28 for the random forest, meaning confident predictions that were sometimes wrong. Where a school intends to rank students by risk and support a fixed number of them, calibration matters more than the accuracy figure, yet reliability diagrams appear in almost none of the reviewed work.

Encoding and preprocessing choices. Label encoding imposes an artificial ordering on nominal attributes such as parental occupation. Tree learners absorb this at the cost of extra splits, but the effect on attribution rankings is unquantified in the literature.

B. Research Gaps

Table V records the gaps that follow from the review, each paired with the evidence in this paper that establishes it.

Table V. Research gaps identified by the review and the supporting evidence.

Gap	Evidence
Ensemble superiority asserted without inference	No pairwise difference significant; Friedman p = 0.145 (Section VI-C)
Single-partition reporting	Accuracy spans 0.8228 to 0.9620 across ten partitions; ranking reverses (Section VI-D)
Accuracy conflated with usefulness	22.79-point loss when interim grades are withdrawn (Section VI-E)
Complexity assumed necessary	Logistic regression matches stacking on all thresholded metrics (Section VI-B)
Attribution from a single model	Concentration on the leading attribute varies from 15.5% to 76.7% across learners (Section VI-F)
Fairness unexamined	Demographic attributes used as predictors with no audit reported in any reviewed study

The fairness gap is the most serious and the least discussed. The corpora used throughout this literature contain sex, address type, parental occupation, parental education and family cohabitation status. A model trained on these and used to allocate remedial support could direct or withhold it along lines correlated with social background. None of the twenty-nine primary studies reviewed reports an equalised-odds or demographic-parity analysis. A related concern is the self-fulfilling character of risk labelling: a student identified as likely to fail may be treated differently in ways that affect the outcome.

8 FUTURE RESEARCH DIRECTIONS

Six directions follow directly from the gaps above, ordered by the ratio of expected benefit to effort.

a) Timely behavioural instrumentation. The S3 collapse shows that the attributes schools currently record cannot support early prediction. Weekly attendance, homework submission, formative assessment marks and learning-platform interaction are the natural substitutes, and the learning-analytics literature shows that derived engagement measures outperform raw activity counts [58], [59]. This is a data-collection problem, not a modelling problem, and it is the binding constraint on the field.

b) Reporting standards. Multi-partition reporting with dispersion, paired significance testing and an explicit statement of the feature scenario should become the minimum for publication. None of the three is expensive; their absence is convention rather than necessity.

c) Fairness auditing and mitigation. Equalised-odds and demographic-parity gaps should be computed across sex, address type and parental education for every model proposed for deployment, with the cost of removing protected attributes reported alongside.

d) Nested cross-validation with tuning. Comparisons should place hyperparameter search in an inner loop and evaluation in an outer loop so that tuning gains are obtained without selection optimism; random search is preferable to grid search at these sample sizes [60].

ISSN: **3139-0862**

e) Cost-sensitive and threshold-aware formulations. The consequences of the two error types differ sharply in this application. Replacing the symmetric loss with an asymmetric one, and reporting the mapping from decision threshold to the number of students flagged, would make outputs directly usable by a school with fixed support capacity.
f) Prospective evaluation. Every result in this literature, including those reported here, is retrospective. Whether acting on these predictions improves outcomes is untested and requires a design in which a flagged group receives support and a comparable group does not.

9 CONCLUSION

This review surveyed twenty-nine primary studies on ensemble methods for student performance prediction and grounded the survey in a controlled experiment on a benchmark widely used within it. Three conclusions follow.

The first is that the evidence for ensemble superiority is weaker than the literature's confidence implies. Ensembles produced the best point estimates in the case study, but no pairwise difference reached significance, the ranking reversed under resampling, and a logistic regression matched the strongest ensemble on every thresholded metric. The reasonable position is that heterogeneous combination gives a small, consistent and statistically unverified advantage on cohorts of this size, and that reporting it as established superiority is not supported.

The second is that reported accuracy and operational usefulness are different quantities, and the literature routinely conflates them. Withdrawing assessment results from the target course cost the best model 22.79 percentage points and reduced discrimination to near chance. Published accuracies on such corpora should always be reported alongside the feature scenario, and comparisons across studies that do not state the scenario are not interpretable.

The third is where the obstacle actually lies. It is not the choice between bagging and boosting, nor the tuning of a learning rate. It is the absence of timely behavioural data in the records schools currently keep. Ensemble methods will be ready when that data arrives; the useful contribution of a review at this point is to say clearly what they can and cannot do with the data that exists now.

Recommendations. Practitioners should predict after the first assessment period rather than before it, select the decision threshold from available support capacity rather than convention, prefer the cheaper model where ranking quality is not required, and deliver an explanation with every prediction. Researchers should report feature scenarios explicitly, test differences before claiming them, repeat over partitions, and audit for fairness before proposing deployment.

REFERENCES

[1] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Trans. Syst., Man, Cybern. C, vol. 40, no. 6, pp. 601-618, Nov. 2010. doi: 10.1109/TSMCC.2010.2053532.
[2] C. Romero and S. Ventura, "Data mining in education," WIREs Data Mining Knowl. Discov., vol. 3, no. 1, pp. 12-27, Jan. 2013. doi: 10.1002/widm.1075.
[3] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," WIREs Data Mining Knowl. Discov., vol. 10, no. 3, art. e1355, May 2020. doi: 10.1002/widm.1355.
[4] T. G. Dietterich, "Ensemble methods in machine learning," in Multiple Classifier Systems (Lecture Notes in Computer Science, vol. 1857), Berlin, Germany: Springer, 2000, pp. 1-15. doi: 10.1007/3-540-45014-9_1.
[5] L. Breiman, "Bagging predictors," Mach. Learn., vol. 24, no. 2, pp. 123-140, Aug. 1996. doi: 10.1007/BF00058655.
[6] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, Oct. 2001. doi: 10.1023/A:1010933404324.
[7] R. E. Schapire, "The strength of weak learnability," Mach. Learn., vol. 5, no. 2, pp. 197-227, Jun. 1990. doi: 10.1007/BF00116037.
[8] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," J. Comput. Syst. Sci., vol. 55, no. 1, pp. 119-139, Aug. 1997. doi: 10.1006/jcss.1997.1504.
[9] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Statist., vol. 29, no. 5, pp. 1189-1232, Oct. 2001. doi: 10.1214/aos/1013203451.
[10] D. H. Wolpert, "Stacked generalization," Neural Netw., vol. 5, no. 2, pp. 241-259, 1992. doi: 10.1016/S0893-6080(05)80023-1.
[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, San Francisco, CA, USA, Aug. 2016, pp. 785-794. doi: 10.1145/2939672.2939785.

[12] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems, vol. 30, Long Beach, CA, USA, Dec. 2017, pp. 3146-3154. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[13] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," in Proc. 5th Annu. Future Business Technology Conf. (FUBUTEC), Porto, Portugal, Apr. 2008, pp. 5-12. [Online]. Available: http://hdl.handle.net/1822/8024

[14] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Comput., vol. 10, no. 7, pp. 1895-1923, Oct. 1998. doi: 10.1162/089976698300017197.

[15] Demsar, "Statistical comparisons of classifiers over multiple data sets," J. Mach. Learn. Res., vol. 7, pp. 1-30, Jan. 2006. [Online]. Available: https://jmlr.org/papers/v7/demsar06a.html

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, Long Beach, CA, USA, Dec. 2017, pp. 4765-4774. [Online]. Available: https://arxiv.org/abs/1705.07874

[17] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nat. Mach. Intell., vol. 2, no. 1, pp. 56-67, Jan. 2020. doi: 10.1038/s42256-019-0138-9.

[18] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, no. 5, pp. 206-215, May 2019. doi: 10.1038/s42256-019-0048-x.

[19] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009. doi: 10.1007/978-0-387-84858-7.

[20] T. K. Ho, "The random subspace method for constructing decision forests," IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 8, pp. 832-844, Aug. 1998. doi: 10.1109/34.709601.

[21] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Mach. Learn., vol. 63, no. 1, pp. 3-42, Apr. 2006. doi: 10.1007/s10994-006-6226-1.

[22] Z. Fan, J. Gou, and S. Weng, "Predicting secondary school student performance using a double particle swarm optimization-based categorical boosting model," Eng. Appl. Artif. Intell., vol. 124, art. 106649, Sep. 2023. doi: 10.1016/j.engappai.2023.106649.

[23] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," Mach. Learn., vol. 51, no. 2, pp. 181-207, May 2003. doi: 10.1023/A:1022859003006.

[24] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms. Boca Raton, FL, USA: CRC Press, 2012. doi: 10.1201/b12207.

[25] O. Sagi and L. Rokach, "Ensemble learning: A survey," WIREs Data Mining Knowl. Discov., vol. 8, no. 4, art. e1249, Jul. 2018. doi: 10.1002/widm.1249.

[26] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch, "Predicting student performance: An application of data mining methods with an educational web-based system," in Proc. 33rd Annu. Frontiers in Education Conf. (FIE), Westminster, CO, USA, Nov. 2003, pp. T2A13-T2A18. doi: 10.1109/FIE.2003.1263284.

[27] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Preventing student dropout in distance learning using machine learning techniques," Appl. Artif. Intell., vol. 18, no. 5, pp. 411-426, May 2004. doi: 10.1080/08839510490442058.

[28] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, "Models for early prediction of at-risk students in a course using standards-based grading," Comput. Educ., vol. 103, pp. 1-15, Dec. 2016. doi: 10.1016/j.compedu.2016.09.005.

[29] J. Xu, K. H. Moon, and M. van der Schaar, "A machine learning approach for tracking and predicting student performance in degree programs," IEEE J. Sel. Topics Signal Process., vol. 11, no. 5, pp. 742-753, Aug. 2017. doi: 10.1109/JSTSP.2017.2692560.

[30] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," Comput. Educ., vol. 113, pp. 177-194, Oct. 2017. doi: 10.1016/j.compedu.2017.05.007.

[31] C. Zabriskie, J. Yang, S. DeVore, and J. Stewart, "Using machine learning to predict physics course outcomes," Phys. Rev. Phys. Educ. Res., vol. 15, no. 2, art. 020120, Oct. 2019. doi: 10.1103/PhysRevPhysEducRes.15.020120.

[32] M. Yagci, "Educational data mining: Prediction of students' academic performance using machine learning algorithms," Smart Learn. Environ., vol. 9, art. 11, Mar. 2022. doi: 10.1186/s40561-022-00192-z.

[33] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Multi-split optimized bagging ensemble model selection for multi-class educational data mining," Appl. Intell., vol. 50, no. 12, pp. 4506-4528, Dec. 2020. doi: 10.1007/s10489-020-01776-3.

[34] A. Moubayed, M. Injadat, A. B. Nassif, H. Lutfiyya, and A. Shami, "E-learning: Challenges and research opportunities using machine learning and data analytics," IEEE Access, vol. 6, pp. 39117-39138, 2018. doi: 10.1109/ACCESS.2018.2851790.

[35] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," Interact. Learn. Environ., vol. 31, no. 6, pp. 3360-3379, 2023. doi: 10.1080/10494820.2021.1928235.

[36] D. Duan, C. Dai, and R. Tu, "Research on the prediction of students' academic performance based on XGBoost," in Proc. 10th Int. Conf. Educational Innovation Through Technology (EITT), Chongqing, China, Dec. 2021, pp. 316-319. doi: 10.1109/EITT53287.2021.00068.

[37] Y. Chen and D. Zhai, "A comparative study on student performance prediction using machine learning," Educ. Inf. Technol., vol. 28, no. 9, pp. 12039-12057, Sep. 2023. doi: 10.1007/s10639-023-11672-1.

[38] W. Hoyos and I. Caicedo-Castro, "Tuning data mining models to predict secondary school academic performance," Data, vol. 9, no. 7, art. 86, Jul. 2024. doi: 10.3390/data9070086.

[39] H. Karalar, C. Kapucu, and H. Guruler, "Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system," Int. J. Educ. Technol. High. Educ., vol. 18, art. 63, Dec. 2021. doi: 10.1186/s41239-021-00300-y.

[40] S. B. Keser and S. Aghalarova, "HELA: A novel hybrid ensemble learning algorithm for predicting academic performance of students," Educ. Inf. Technol., vol. 27, no. 4, pp. 4521-4552, May 2022. doi: 10.1007/s10639-021-10780-0.

[41] N. A. Butt, Z. Mahmood, K. Shakeel, S. Alfarhood, M. Safran, and I. Ashraf, "Performance prediction of students in higher education using multi-model ensemble approach," IEEE Access, vol. 11, pp. 136091-136108, 2023. doi: 10.1109/ACCESS.2023.3336987.

[42] S. A. Priyambada, T. Usagawa, and M. Er, "Two-layer ensemble prediction of students' performance using learning behavior and domain knowledge," Comput. Educ. Artif. Intell., vol. 5, art. 100149, 2023. doi: 10.1016/j.caeai.2023.100149.

[43] E. Cetin and O. Cetin, "Stacking: An ensemble learning approach to predict student performance in PISA 2022," Educ. Inf. Technol., vol. 30, pp. 4235-4259, 2025. doi: 10.1007/s10639-024-13110-2.

[44] M. Adnan et al., "Predicting at-risk students at different percentages of course length for early intervention using machine learning models," IEEE Access, vol. 9, pp. 7519-7539, 2021. doi: 10.1109/ACCESS.2021.3049446.

[45] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics dataset," Sci. Data, vol. 4, art. 170171, Nov. 2017. doi: 10.1038/sdata.2017.171.

[46] K. Talebi, Z. Torabi, and N. Daneshpour, "Ensemble models based on CNN and LSTM for dropout prediction in MOOC," Expert Syst. Appl., vol. 235, art. 121187, Jan. 2024. doi: 10.1016/j.eswa.2023.121187.

[47] Q. Liu et al., "EKT: Exercise-aware knowledge tracing for student performance prediction," IEEE Trans. Knowl. Data Eng., vol. 33, no. 1, pp. 100-115, Jan. 2021. doi: 10.1109/TKDE.2019.2924374.

[48] A. Malik et al., "Forecasting students' adaptability in online entrepreneurship education using modified ensemble machine learning model," Array, vol. 19, art. 100303, Sep. 2023. doi: 10.1016/j.array.2023.100303.

[49] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed. Munich, Germany: Independently published, 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[50] S. Wang and B. Luo, "Academic achievement prediction in higher education through interpretable modeling," PLoS ONE, vol. 19, no. 9, art. e0309838, Sep. 2024. doi: 10.1371/journal.pone.0309838.

[51] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, art. 6, Jan. 2020. doi: 10.1186/s12864-019-6413-7.

ISSN: **3139-0862**

[52] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS ONE, vol. 10, no. 3, art. e0118432, Mar. 2015. doi: 10.1371/journal.pone.0118432.

[53] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321-357, Jun. 2002. doi: 10.1613/jair.953.

[54] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153-157, Jun. 1947. doi: 10.1007/BF02295996.

[55] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," J. Amer. Statist. Assoc., vol. 32, no. 200, pp. 675-701, Dec. 1937. doi: 10.1080/01621459.1937.10503522.

[56] [56] P. Cortez, "Student Performance," UCI Machine Learning Repository, 2014. [Online]. Available: https://archive.ics.uci.edu/dataset/320/student+performance

[57] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825-2830, Nov. 2011. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[58] G. Akcapinar, A. Altun, and P. Askar, "Using learning analytics to develop early-warning system for at-risk students," Int. J. Educ. Technol. High. Educ., vol. 16, art. 40, Nov. 2019. doi: 10.1186/s41239-019-0172-z.

[59] L. A. B. Macarini, C. Cechinel, M. F. B. Machado, V. F. C. Ramos, and R. Munoz, "Predicting students success in blended learning: Evaluating different interactions inside learning management systems," Appl. Sci., vol. 9, no. 24, art. 5523, Dec. 2019. doi: 10.3390/app9245523.

[60] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," J. Mach. Learn. Res., vol. 13, pp. 281-305, Feb. 2012. [Online]. Available: https://jmlr.org/papers/v13/bergstra12a.html