A Multi-Layer Computational Framework for Predicting Student Performance Ranges in Higher Education Using Machine Learning

Abdellatif HARIF, Moulay Abdellah KASSIMI
Laboratory of Science of Information Technology Data-Mathematics and Applications National School of Applied Sciences, IBN ZOHR University, Agadir, Morocco

***Abstract*—Predicting student academic performance constitutes a strategic priority for higher education institutions seeking to reduce attainment gaps and provide timely, targeted support. Existing approaches predominantly generate single-point performance estimates, overlooking the inherent variability in individual academic trajectories. This paper introduces a novel seven-layer computational framework that predicts student performance as a bounded range, capturing both minimum and maximum expected outcomes rather than as a solitary value. The framework integrates a bespoke imbalanced-data mitigation algorithm, three heuristic feature-selection methods: Genetic Algorithm, Particle Swarm Optimization, and Recursive Feature Elimination, and two complementary model architectures: a Parallel Architecture built upon fourteen supervised learning classifiers, and a Popularity Architecture centered on K-Modes/K-Prototype unsupervised clustering. The framework was validated on a rich, anonymized dataset provided by IBN ZOHR University in Morocco, comprising records from over 200,055 undergraduate students. The proposed framework achieves accuracy of 84%/86% (worst/common-case scenario), representing a 3%/5% improvement over an 81% baseline derived from the ten most relevant prior studies. The unsupervised Popularity Architecture attained peak accuracy of 96.91%, outperforming all supervised configurations. Results further demonstrate that omitting feature selection frequently yields competitive performance, and that increasing the number of hidden layers in neural networks does not significantly alter predictive accuracy in this educational context. The framework is designed for seamless integration into existing student performance dashboard systems, offering the institutions an actionable decision-support tool.**

Keywords*—Student performance prediction; machine learning; unsupervised learning; performance range; higher education; educational data mining; feature selection

I. Introduction

Monitoring and predicting academic performance have become a cornerstone of institutional strategy in higher education. Proactively identifying students who are at risk of underperforming, or conversely those likely to excel, enables universities to deploy targeted support mechanisms, optimize resource allocation, and reduce systemic attainment disparities. With the proliferation of administrative student information systems, educational institutions now accumulate rich repositories of longitudinal data that span demographic profiles, enrollment characteristics, attendance records, and progression trajectories. These repositories present a compelling opportunity to develop data-driven decision support tools grounded in machine learning (ML).

Research at the intersection of Artificial Intelligence (AI) and Education has been active since the early 1990s, when logistic regression models were first applied to predict graduate retention [1]. Over the subsequent three decades, the field expanded to encompass a broad repertoire of supervised learning algorithms—Random Forest, Support Vector Machines, Naive Bayes, K-Nearest Neighbors, and Gradient Boosting variants—alongside deep architectures such as Feed Forward Neural Networks (FFNNs) [2], [3], [4]. Despite this proliferation, two structural limitations persist across the literature.

First, virtually all published studies produce predictions as singular values—either a numerical score or a discrete classification label—thereby ignoring the natural performance variability that students exhibit across assessments, life circumstances, and academic cycles. A student's academic trajectory is rarely deterministic; financial pressures, health events, and personal commitments introduce stochastic fluctuations that a point estimate cannot capture. Second, the issue of class imbalance in educational datasets, most pronounced in demographic variables, remains inadequately addressed leaving predictive models susceptible to biased outcomes that favor the majority class [5].

This paper addresses both constraints through a comprehensive and reproducible computational framework. Its main novelty lies in predicting academic performance as a defined range, expressed as lower and upper values of the consistency scale, rather than a single predicted score. This range reveals the limits of a student's potential academic ability under varying conditions, providing more accurate and practically useful outputs for decision-makers in educational institutions.

This framework also contributes and allowing: First, a native, proven data training algorithm to reduce the likelihood of data imbalance and associated overshoot risks; Second, a popular architecture that incorporates unsupervised clustering as a preliminary step to performance prediction, an approach that has not been widely explored in the literature; Third, a systematic comparison of eight feedforward neural network architectures with varying depths of hidden layers; Fourth, an in-depth multi-method data analysis layer that includes principal

component analysis (PCA), Multiple Correspondence Analysis (MCA), factor analysis of mixed data (FAMD), and statistical hypothesis testing; and a custom-designed scaling-of-validity measure compatible with scaling-value predictions.

The remainder of this article is divided as follows: Section II reviews the existing literature and places it in its critical context. Section III presents the proposed computational framework, and Section IV describes experimental verification. Section V presents and discusses the results, and Section VI concludes with a summary of the contributions and guidelines for future research.

II. Literature Review

A. Scope and Methodology

A systematic literature review was conducted to map published research on predicting student performance using artificial intelligence. Before examining each study individually, several comprehensive meta-analyses were consulted to gain a broader understanding of the field. Notably, Juho Hellas et al. [6] synthesized 357 studies published between 2010 and 2018, while Faisal Alwarthan et al. [5] analyzed 44 studies covering the period from 2010 to 2022. Similarly, Muhammad Chaudhry and Muhammad Kazim [7] reviewed 85 studies published between 2011 and 2021, and Virginia Dignum [8] examined 35 works spanning from 2000 to 2020.

More recent contributions have further expanded this body of knowledge. In particular, Batool et al. [9] surveyed approximately 260 studies published between 2000 and 2022, providing one of the most comprehensive overviews to date. In addition, Alnasyan et al. [10] focused specifically on deep learning approaches, reviewing 46 studies published between 2019 and 2023.

The primary literature search was conducted through Google Scholar, IEEE Xplore, ScienceDirect, arXiv, and ResearchGate, applying conjunctive keyword combinations such as 'predict student performance AND Supervised Learning', 'academic prediction AND Neural Networks', and 'student dropout AND Machine Learning'. Papers were included if they were in English, reported sufficient methodological detail for reproducibility, explicitly addressed performance prediction with AI methods, and had received at least one citation.

Together, these reviews form the baseline against which the present work is positioned.

B. Supervised Learning Approaches

1) Regression and logistic regression: Regression-based models constitute one of the earliest and most persistent branches of this literature. Pyke and Sheridan [1] demonstrated that logistic regression could predict graduate student retention with accuracy in the range of 77%–88%, drawing on demographic and funding variables. Yang et al. [11] showed that combining multiple linear regression with pca-based dimensionality reduction yielded a 50% improvement in predictive error compared to regression alone, attributable to the removal of multicollinear features. Pereira [12] enriched this paradigm by pairing linear regression with FAMD, achieving accuracy of approximately 78% on a 1,044 student Portuguese secondary school dataset. These works collectively suggest that dimensionality reduction is a valuable complement to regression when datasets contain predominantly qualitative variables.

Also, several studies focus on specific institutional examination contexts. Utzman, Riddle, and Jewell [13] applied logistic regression to predict performance on the National Physical Therapy Examination, finding that prior academic qualifications, race, and ethnicity were the most informative predictors, though without evidence of a direct causal relationship. Jaber et al. [14] explored the predictive power of weekly quizzes for medical students' shelf exam performance, confirming that later-week assessments are stronger predictors than earlier ones.

2) Ensemble and other supervised learning models: Ensemble methods—particularly random forest and its variants—have emerged as dominant performers in educational prediction. Adekitan and Noma-Osaghae [2] demonstrated their superiority over classical models on first-year engineering student data. Mengash [15] integrated random forest and decision tree classifiers within a multi-state university admissions framework, reporting accuracy rates consistent with institutional requirements. Authors in [16] compared four classifiers (Naive Bayes, Decision Tree, Random Forest, SVM) and found ensemble methods most stable across cross-validation folds.

Gradient Boosting variants remain underexplored in this specific domain, with only a handful of studies applying XGBoost [17] and fewer applying Extra Trees. This gap motivates the broader model comparison conducted in the present study. Notably, [3] is one of the few studies applying seven different classifiers in a single experiment, achieving accuracy of 71%–82%—the largest model comparison identified in the literature prior to this work.

3) Feed forward neural networks: FFNN-based approaches have attracted growing attention, particularly for their capacity to detect non-linear relationships in high-dimensional data. Bilal et al. [18] demonstrated that a single-hidden-layer network trained on enrollment data achieves accuracy exceeding 90%. The author in [19] reported 95% accuracy using a four-variable ANN on moodle learning management data, though the small dataset (n = 265) raises generalization concerns. waheed et al. [4] applied deep neural networks to Open University Learning Analytics data, achieving improved accuracy compared to shallower architectures, it was one of the rare studies using more than two hidden layers.

Despite these successes, literature consistently employs a single FFNN architecture per study, typically with one or two hidden layers and an arbitrarily chosen node count. Whether additional hidden layers systematically improve accuracy in the educational prediction context remains an open question that the present study explicitly investigates.

C. Unsupervised Learning

The application of unsupervised learning to student performance prediction is exceedingly rare. Alwarthan et al. [5] identified only one study employing clustering as part of a prediction pipeline. Ding et al. [20] employed unsupervised feature learning within an MOOC prediction context, demonstrating that unsupervised representations improved the downstream supervised classifier. The present work extends this precedent by proposing a fully articulated Popularity Architecture in which clustering directly produces the performance-range prediction, by passing the need for labelled training targets during the prediction phase.

Several additional gaps are evident from the reviewed literature. Imbalanced data is acknowledged as a problem but rarely addressed with purpose-designed algorithms—most studies either ignore it or apply generic dimensionality reduction. In-depth statistical data analysis is largely absent; most papers report only basic descriptive statistics and correlation matrices. Finally, the output format invariably takes the form of a single value, leaving the inherent performance variability of students unmodelled.

D. Summary and Research Gaps

The reviewed literature reveals five principal gaps that the present work addresses. First, there is a predominance of single-value performance predictions that ignore performance variability. Second, there is insufficient treatment of imbalanced data. Third, the limited use of unsupervised learning. Fourth, shallow data analysis layers that fail to expose dataset strengths and weaknesses. Fifth is the absence of systematic FFNN architecture comparisons within a single experiment. Table I provides a comparative overview of the studies most closely related to the proposed framework.

TABLE I. COMPARATIVE OVERVIEW OF CLOSELY RELATED STUDIES. SL = SUPERVISED LEARNING ; NN = NEURAL NETWORK; UL = UNSUPERVISED LEARNING

Study	Algorithms	Number of Models	Output Format	Imbalance Addressed
Bilal et al. [18]	NN	1	Value	No
Rodríguez-Hernández et al. [3]	SL + NN	7	Value	No
Helal et al. [21]	SL + NN	4	Value	No
Mengash [15]	SL + NN	4	Value	No
Waheed et al. [4]	NN (Deep)	1	Value	No
Zhao et al. [22]	SL + NN	8	Value	No
Francis & Babu [23]	SL + NN	4	Value	No
Aissaoui et al. [24]	SL	1	Value	No
Bettahi et al. [25]	SL	3+	Value	Yes
Fazil et al. [26]	NN (Deep)	1	Value	No
Proposed Framework	**SL + NN + UL**	**18+**	**Range**	**Yes**

III. COMPUTATIONAL FRAMEWORK

The proposed computational framework is organized as a sequential pipeline of seven layers, each with a defined input specification, processing function, and output contract. The layers are executed in strict chronological order to ensure that downstream components receive validated, properly formatted inputs. Fig. 1 schematizes the full pipeline. The following subsections describe each layer in detail.

A. Data Pre-Processing and Exploratory Analysis

Raw educational datasets are typically stored at the module level, with multiple rows per student. This layer restructures the data into a single student-centric representation through five operations: 1) grouping records by student identifier, averaging quantitative variables, and retaining the first-occurrence value for qualitative ones; 2) removing incomplete rows to avoid imputation bias; 3) discarding variables that become uninformative after aggregation; 4) engineering four features: student age, number of modules completed, geographic distance to the faculty, and the Consistency-Scale, a discretized grade metric mapped to a 0–1 ordinal scale in 0.1 increments; finally 5) retaining only the descriptive version of duplicated variable definitions. Once pre-processing is complete, an exploratory analysis is conducted to assess dataset quality, variable distributions, and inter-variable relationships. The protocol applies frequency distributions and descriptive statistics for qualitative and quantitative variables respectively, alongside Pearson correlation matrices, normality tests, dimensionality reduction, and stratified analyses by performance group.

B. Prediction Pre-Processing

The second layer configures the data for machine learning. Three transformations are applied: 1) selection of the target student subgroup—in the present experiments, only graduate students who completed their degree program are retained, and only enrollment-stage variables are included, reflecting the intention to support early prediction at the point of admission; 2) label encoding of all qualitative variables to convert categorical attributes into integer-valued identifiers. Finally, 3) Z-score normalization of quantitative variables and ordinal scaling of qualitative variables to equalize feature magnitudes and improve model convergence speed.

C. Feature Selection

The third layer explores which subsets of input features are most informative for predicting the Consistency-Scale target. Three meta-heuristic feature-selection algorithms are applied independently: 1) the Genetic Algorithm (GA), which evolves a population of candidate feature subsets over multiple generations, selecting subsets that maximize a cross-validated accuracy fitness function; 2) Particle Swarm Optimization (PSO), which positions 50 particles in feature-space, each representing a candidate subset, and iteratively refines their positions by learning from the globally best-performing particle; and 3) Recursive Feature Elimination (RFE), which fits a logistic regression model, ranks features by coefficient magnitude, and iteratively discards the lowest-ranked feature until a minimum subset size is reached. A fourth experimental condition retains all features without selection, serving as a baseline. For each algorithm, the internal evaluation model is

logistic regression, and a 5-fold cross-validation scheme is applied.

D. Training/Testing Split

The fourth layer in our framework, partitions the student-level dataset D into a training D_train (80%) and a test set D_test (20%). Data is randomly shuffled prior to splitting to prevent ordering artefacts. The 80:20 ratio was selected as a standard practice that provides sufficient data for model training while retaining an adequate holdout for unbiased evaluation.

E. Anchored Training Data

As illustrated in Fig. 2, the fifth layer addresses the problem of imbalanced class distributions in educational data by identifying an optimal training sample size. The Anchored Training Data method(ADT), formally described in Algorithm 1, operates as follows: D_train is divided into five equal-sized segments s_1, …, s_5. In iteration i (i = 1, …, 5), the concatenation of segments s_1 through s_i forms a temporary training set D_s, which is further split 80:20 and used to train a logistic regression classifier. The segment count whose corresponding accuracy benchmark is highest is selected as the optimal training size. This progressive segment accumulation mirrors the intuition of walk-forward validation while specifically targeting the imbalance reduction property: a smaller, more carefully selected training partition tends to exhibit more balanced class ratios than the full dataset, thereby reducing overfitting risk.

Algorithm 1: Anchored Training Data (ATD)

Input: Training sample set D_train (shuffled); S = 5 segments
1: Initialize the full segment set S = {s1, s2, …, s5} by randomly partitioning D_train into 5 equal-sized disjoint segments
2: Initialize best_acc ← 0 ; best_i ← 1
3: Specify n_folds = 5 for the internal cross-validation split
4: While i ⩽ S do
5: For (each accumulated subset) do
6: D_s ← concatenate(s1, s2, ⋯, si) // Progressive accumulation
7: Randomly split D_s into D_s_train (80%) and D_s_val (20%)
8: Train logistic regression M_i on D_s_train
9: Perform ValidRangeAccuracy(M_i, D_s_val) → acc_i
10: Rank acc_i and compare against best_acc
11: If (acc_i > best_acc) then
12: best_acc ← acc_i
13: best_i ← i
14: End If
15: End For
16: i ← i + 1
17: End While
18: D_optimal ← concatenate(s1, ⋯, s{best_i}) using the segment count with best_acc
19: Output D_optimal as the selected training subset for downstream model training

F. Model Architectures

The sixth layer implements predictions using two complementary architectures.

1) Parallel architecture (supervised learning): The Parallel Architecture employs supervised classifiers to predict the minimum and maximum Consistency-Scale values independently for each student in D_test. Sixteen classifiers are applied: Random Forest (RFC), Adaboost (ADAC), K-Nearest Neighbours (KNNC), Gradient Boost (GBC), Gaussian Process (GPC), Passive Aggressive (PAC), Extra Trees (ETC), Extreme Gradient Boost (XGBC), and eight FFNN configurations crossing four depth levels (1, 3, 5, 10 hidden layers) with two node counts (4, 8 nodes per layer). each classifier is trained separately on the minimum and maximum consistency-scale targets, yielding two predictions per student that together constitute the predicted performance range.

2) Popularity architecture (unsupervised learning): The Popularity Architecture replaces direct supervised prediction with a cluster-based inference mechanism, in which unsupervised grouping of enrollment profiles substitutes for labelled target learning. The generalization process operates in two phases. The optimal number of clusters k is first determined using the Davies–Bouldin index, preferred over the Calinski–Harabasz score for its compatibility with non-euclidean settings and qualitative variables. The K-Modes model (purely qualitative features) or K-Prototype (mixed features) is then fitted on D_train, producing k centroids. for each cluster, Consistency-scale values with frequencies above the cluster median are designated as popular, and their minimum and maximum define the cluster-specific bounds [C_min_k, C_max_k]. during inference, each student in D_test is assigned to the cluster whose centroid minimizes the cao dissimilarity, a distance measure compatible with both qualitative and quantitative features. The corresponding bounds [C_min_k, C_max_k] are returned as the predicted performance interval. this inductive mechanism requires no target labels at test time: predictions derive entirely from the training clusters' popularity distributions, while the learned centroids serve as stable reference points for assigning new observations.

G. Valid Range Benchmark

Since the proposed framework produces interval-valued predictions rather than point estimates, standard classification metrics (accuracy, F1, AUC) and regression metrics (MAE, MSE) are not directly applicable. A custom metric — the Valid Range — is therefore introduced. A prediction is classified as correct if and only if the student's true average Consistency-Scale value (C_avg) falls within the predicted interval [C_min, C_max], and the aggregate Valid Range Accuracy is computed as follows:

Valid Range Accuracy = (Number of Correct Predictions / Total Predictions) × 100%.

To address the concern that wide prediction intervals may artificially inflate Valid Range Accuracy, two complementary measures are reported alongside each architecture's coverage score. First, the mean predicted interval width W = mean(C_max − C_min) quantifies interval informativeness: narrower intervals at high coverage indicate stronger predictive specificity. Second, a Midpoint Mean Absolute Error (M-MAE) — defined as the mean absolute difference between the interval midpoint (C_min + C_max)/2 and the true C_avg — provides a scalar error measure compatible with standard regression benchmarks. A

coverage-width efficiency score E = Coverage / W is additionally reported, where higher values reflect better accuracy per unit of interval width. Together, these measures form a comprehensive evaluation protocol for interval-valued predictions.

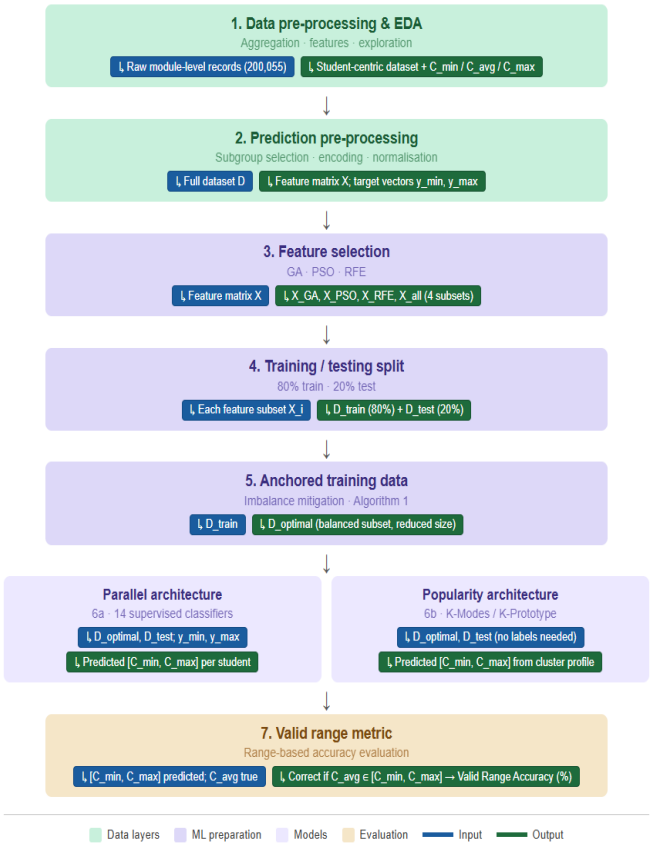

Fig. 1. Seven-layer computational framework for student performance range prediction. layer interactions, inputs/outputs, and derivation of performance bounds [C_min, C_max].

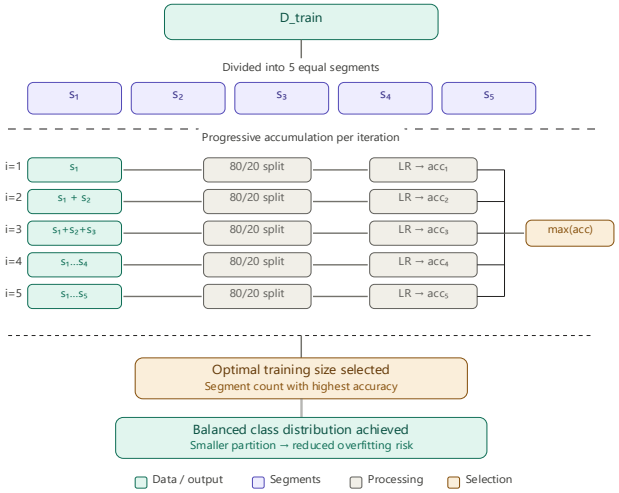

Fig. 2. Anchored training data method: progressive segment accumulation for optimal training size selection.

IV. Experimental Validation

A. Dataset Description

The dataset was provided by Ibn Zohr University of Morocco and covers academic records from 2019 to 2025. It comprises 200,055 records drawn from two institutional sources: a pre-registration platform capturing student-declared information at enrollment — including age, gender, region of residence, and socioeconomic background — and an academic management system centralizing module-level results across each semester and academic year. All records were anonymized prior to use, with no personally identifiable information retained, in compliance with Moroccan data protection legislation.

Table II describes the full variable set. The dataset is mixed-type, containing nominal, ordinal, discrete, and continuous variables. This heterogeneity directly shaped key design decisions in the framework — most notably the adoption of K-Prototype clustering, which is specifically suited to datasets combining qualitative and quantitative inputs.

B. Data Analysis Findings

The data analysis layer revealed several educationally significant patterns. Quantitative analysis using Pearson's correlation coefficient, as shown in Fig. 3, demonstrated moderate positive correlations between module score and Presence Rate (r = 0.388), and between module score and Presence Duration (r = 0.038), indicating that punctuality is associated with better performance. Geographic proximity to the university institution (Lodging distance) also showed a weak but positive correlation with scores, consistent with previous findings on the relationship between proximity and academic results [27].

Fig. 3. Pearson Correlation Matrix of Quantitative Variables

C. Feature Selection Results

Table III summarizes the features selected by each algorithm. Lives Environment, Parents Relationship, and Student with certificates were each selected by all three algorithms, reflecting their consistent predictive relevance. Home Distance was selected by the GA but not PSO or RFE, while Module Score was selected by RFE alone indicating some

divergence in the features prioritized by each method. The 'no feature selection' condition retains all thirteen candidate features and serves as an important comparison point. This condition frequently produced competitive or superior accuracy, consistent with the theory that feature selection provides the greatest benefit when many uninformative features dilute the signal [28].

TABLE II. OUR DATASET VARIABLE DESCRIPTIONS

Column	Description	Format	Attributes	Type
Student code	The student's unique code.	String	200,055	Nominal
Gender	The student's gender.	String	2	Nominal
Age Enrollment	The student's age at the time of enrollment.	Integer	$17 \leq x \leq 65$	Discrete
Lives Environment	Type of lives environment	String	3	Nominal
Limited mobility	If student is Disabled	string	2	Nominal
Parents Relationship	relationship of student's parent	Integer	2	Ordinal
Student with certificates	Students have other certificates	Integer	3	Ordinal
Socioeconomic Class	Student's Profession	String	2	Ordinal
Mother's Profession	Mother's educational level	String	3	Ordinal
Father profession	father's educational level	String	3	Ordinal
Mother Profession	mother's Profession	String	7	Ordinal
Father Profession	Father's Profession	String	7	Ordinal
Home region	The student's home region.	String	4	Nominal
Travel Type	The student's method of traveling to the institution.	String	3	Nominal
Accommodation	The student's study term residence name.	String	8	Nominal
Lodging Distance	The student's accommodation KM distance to the university.	Float	$0 \leq x \leq 300$	Continuous
Home Distance	The student's home KM distance to the university.	Float	$0 \leq x \leq 1500$	Continuous
Academic Years	The student's total number of years of study.	String	4	Ordinal
Status	The student's current academic circumstance.	String	15	Nominal
Specialty chosen	specialty chosen	String	3	Nominal
Faculty name	The faculty the student studied.	String	7	Nominal
Baccalaureate Type	Baccalaureate Type	String	3	Ordinal
Modules (Amount)	The student's total number of modules completed.	Integer	$1 \leq x \leq 42$	Discrete
Module Score	The student's average module grade (%).	Float	$0 \leq x \leq 100$	Continuous
Graduate Grade	The student's final classification grade.	String	4	Ordinal
Standard Grade	Whether the student's average module grade is $\geq 60\%$.	Integer	2	Ordinal
Consistency-Scale	The student's performance scale (min, average & max versions).	Float	10	Ordinal
Course Presence Rate	The student's average attendance rate.	Float	$0 \leq x \leq 1$	Continuous
Lesson presence duration	The student's average presence time per scheduled lesson.	Float	$-29 \leq x \leq 263$	Continuous

TABLE III. FEATURE SELECTION RESULTS ACROSS THREE HEURISTIC ALGORITHMS. TICK INDICATES FEATURE SELECTED.

Features	GA	PSO	RFE	Total Selections
Lives Environment	✓	✓	✓	3
Parents Relationship	✓	✓	✓	3
Student with certificates	✓	✓	✓	3
Travel Type	✓	✓	✓	3
Lodging Distance	✓	✓	✓	3
Age Enrollment	✓	—	✓	2
Father Profession	✓	✓	—	2
Socioeconomic Class	✓	—	✓	2
Accommodation	✓	—	✓	2
Baccalaureate Type	—	✓	✓	2
Gender	—	✓	✓	2
Home Distance	✓	—	—	1
Module Score	—	—	✓	1

D. Model Configuration

For Parallel Architecture, classifiers were configured with standard hyperparameters appropriate for multi-class classification. Random Forest used 30 trees with Gini criterion; Gradient Boost used 100 boosting stages with a learning rate of 0.1; Extreme Gradient Boost used 1,000 iterations, a maximum depth of 5, L1 regularization of 0.3, and L2 of 0.5; K-Nearest Neighbors used k = 5 with Euclidean distance; all eight FFNN variants used 1,000-sample batches and 300 training epochs with a single output node. For Popularity Architecture, the K-Modes/K-Prototype model used 100 iterations with the Cao initialization method; the optimal cluster count was determined separately for each feature-selection condition using the Davies-Bouldin index. Table IV illustrates the configuration used.

TABLE IV. OPTIMAL CLUSTER CONFIGURATIONS FOR THE POPULARITY ARCHITECTURE

Feature Selection	Optimal Clusters (k)	Davies-Bouldin Score
Genetic Algorithm	11	24.77
Particle Swarm Optimisation	5	14.27
Recursive Feature Elimination	7	17.59
No Feature Selection	5	12.48

V. RESULTS AND DISCUSSION

A. Baseline and Comparative Framework

Comparing our results directly against prior work is not straightforward, since the proposed framework predicts performance as a range rather than a single value. We therefore base the comparison on accuracy figures reported under comparable experimental conditions, which is a common and accepted approach in educational data mining benchmarking.

The ten studies forming our baseline were not selected arbitrarily. We applied three filtering criteria: the study had to focus specifically on predicting student academic performance using supervised or neural network methods; it had to report classification accuracy on a proper held-out test set, not merely on cross-validated folds; and it had to use administrative student data from an undergraduate or graduate institutional context. Studies that did not meet all three criteria were excluded.

For each retained study [2,3,15,18,19,21,22,23,29,30], we took the single best accuracy value reported. When a study presented results across multiple configurations, we selected the highest-performing one — reflecting the standard convention of comparing a new framework against the strongest available prior result, not an average one. Across the ten studies, reported accuracy values range from 70% to 90%, yielding a mean baseline of 81% with a standard deviation of ±5.4 percentage points [31].

B. Parallel Architecture: Supervised Learning Results

Fig. 4 summarizes Valid Range accuracy for the eight traditional supervised classifiers across four feature-selection conditions. Random Forest consistently achieved the highest accuracy (93.01%–93.51%), followed closely by Extra Trees (91.52%–92.75%) and Gaussian Process (88.17%–90.31%). The no-feature-selection condition marginally outperformed the three feature-selection conditions for most classifiers, with RFC reaching its peak of 93.51% without feature selection—supporting the general observation that well-curated datasets benefit less from feature pruning.

Extreme Gradient Boost performed anomalously poorly (30.84%–46.66%), an outcome attributed to the predominantly qualitative nature of the dataset and its susceptibility to noise-induced overfitting in this configuration. Excluding XGBC, the average accuracy across all classifiers and feature-selection conditions is approximately 85%, representing a 4% improvement over the baseline. Including XGBC, the average drops to approximately 80%, marginally below baseline—reflecting the distorting influence of a single poorly suited algorithm. This behavior is consistent with Abdulla et al.[28], who noted that qualitative-variable-heavy problems can impair algorithms whose regularization mechanisms assume a quantitative feature structure.

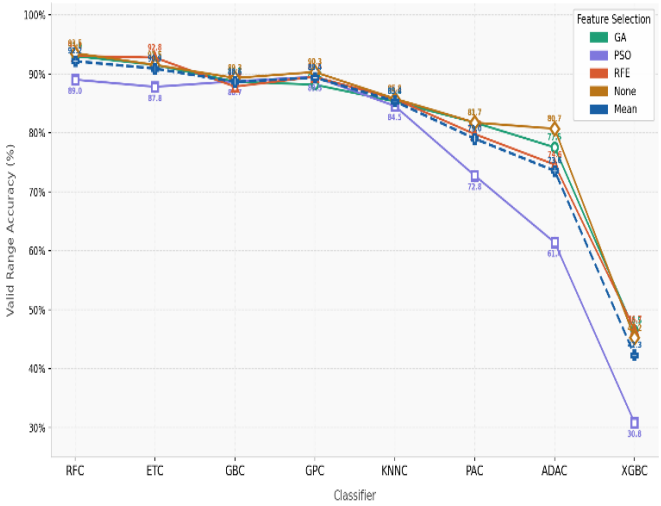

Fig. 4. Valid Range accuracy (%) for supervised classifiers in parallel architecture, by feature-selection condition.

C. Parallel Architecture: Feed Forward Neural Network Results

Fig. 5 reports accuracy for the eight FFNN configurations. PSO-selected features yielded the highest mean accuracy for neural networks (89.20%), followed by no feature selection (86.84%), RFE (87.07%), and GA (86.14%). Peak accuracy of 93.68% was achieved by the NN18 (1 hidden layer, 8 nodes) architecture under PSO feature selection.

Critically, no systematic improvement was observed as the number of hidden layers increased from 1 to 10. The accuracy difference across all FFNN models was approximately 22%, but this variation was attributable to the feature-selection condition rather than network depth. When the worst-performing outlier (NN14 under no feature selection, 72.08%) is excluded as an artefact of a suboptimal node-count and weight initializations combination, the inter-architecture range narrows to approximately 12%. These findings are consistent with prior observations that deeper architectures do not inherently outperform shallower ones in educational datasets of this scale and suggest that the noise level and qualitative composition of the data limit the utility of additional hidden layers [32]. The

overall FFNN mean accuracy of 87% represents a 6% improvement over the baseline.

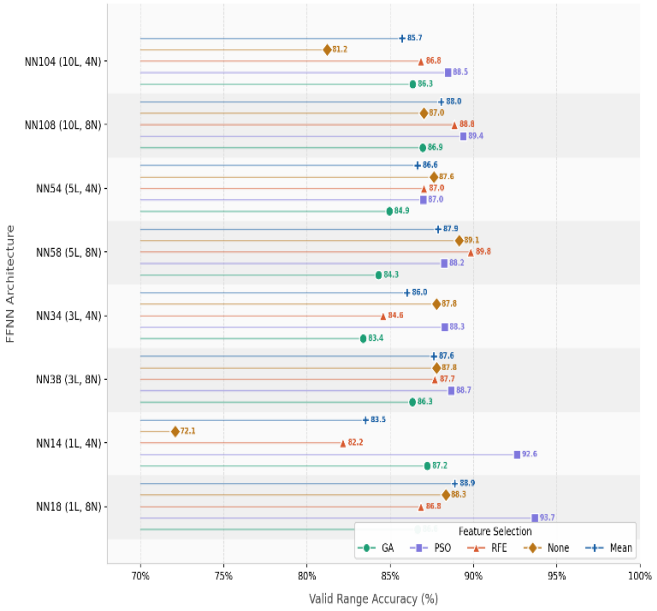

Fig. 5. Valid Range accuracy (%) for FFNN configurations in the Parallel Architecture. L = Hidden Layers; N = Nodes per Layer.

D. Popularity Architecture: Unsupervised Learning Results

The Popularity Architecture — which integrates K-Modes/K-Prototype clustering as the predictive mechanism — achieved the highest coverage across all experimental conditions, as shown in Table V. The no-feature-selection condition produced the best coverage at 96.91%, followed by GA at 95.87%, PSO at 94.10%, and RFE at 92.64%. The coverage range across conditions is narrow (approximately 4%), suggesting that the clustering-based approach is relatively insensitive to the specific feature subset employed — a practically significant property, as it implies that the Popularity Architecture can be deployed without the computational overhead of feature selection.

However, coverage alone does not fully characterize predictive quality. The mean interval width W reveals that the no-feature-selection condition, despite its highest coverage, produces the widest intervals (W = 0.28), shared with PSO. The GA condition, by contrast, achieves a coverage-width efficiency score E = 4.79 — the highest across all configurations — by combining competitive coverage (95.87%) with the narrowest intervals (W = 0.20), a result attributable to the finer cluster granularity of k = 11 compared to k = 5 for the no-selection condition. The M-MAE values (0.040–0.059) confirm that interval midpoints remain close to the true C_avg across all conditions, providing a scalar error measure consistent with standard regression benchmarks.

TABLE V. SUMMARY OF PEAK VALID RANGE ACCURACY RESULTS ACROSS ALL ARCHITECTURES, COMPARED TO THE LITERATURE BASELINE

Architecture	Feature Selection	k (clusters)	Coverage (%)	W (mean width)	M-MAE	Max W (scale)	E = Coverage/W	Diff. vs Baseline (%)
Popularity (UL)	None	5	96.91	0.28	0.053	0.90	3.46	+15.91
Popularity (UL)	GA	11	95.87	0.20	0.040	0.90	4.79	+14.87
Popularity (UL)	PSO	5	94.10	0.28	0.058	0.90	3.36	+13.10
Popularity (UL)	RFE	7	92.64	0.24	0.053	0.90	3.86	+11.64
Parallel (RFC)	None	—	93.51	0.28	0.059	0.90	3.34	+12.51
Parallel (NN18)	PSO	—	93.68	0.26	0.055	0.90	3.60	+12.68
Literature Baseline	—	—	81.00	—	—	—	—	—

The 14%–16% improvement over the supervised baseline warrants careful interpretation. Popularity Architecture benefits from a distinct information structure: rather than learning a mapping from target labels, it groups students by the similarity of their enrollment characteristics and inherits performance bounds from the cluster. This approach is particularly well-suited to the present dataset, where demographic and socioeconomic characteristics appear more informative for group-level stratification than for individual-level regression. The finding is broadly consistent with work by Ding et al. [20], who showed that unsupervised feature representations can boost downstream prediction performance in MOOCs.

E. Overall Analysis and Framework Robustness

Considered holistically, and excluding the anomalous XGBC results, the framework achieves an average accuracy of approximately 86% (range 61%–97%), representing a 5% mean improvement over the 81% baseline (Table V). Including XGBC, the mean drops to 84% (a 3% improvement), with a worst-case accuracy of 31%, reflecting the unsuitability of that specific algorithm for this dataset rather than a systemic weakness of the framework. In both scenarios, the upper-bound accuracy (97%) substantially exceeds the top-end of the baseline range (90%), confirming that the framework can produce state-of-the-art performance even in its worst-case-excluded form.

The robustness of the framework is attributable to the systematic design of its constituent layers. The Anchored Training Data layer reduced training set size from the full 9,300 observations to as few as 3,520 in some conditions, indicating that the algorithm identified smaller, more balanced subsets as preferable training configurations. The in-depth data analysis layer identified key dataset weaknesses—particularly the Lives Environment imbalance and the low dimensionality-reduction fidelity—which informed the decision to restrict the experiment to graduate students with complete enrollment data, thereby maximizing data quality. The multi-algorithm feature selection layer provided evidence that the dataset's predictive power is

broadly distributed across its features, explaining why the no-feature-selection condition frequently performed best (Table V, rows 1 and 5).

To evaluate whether the observed accuracy differences between feature-selection conditions are statistically significant, Wilcoxon signed-rank tests were conducted for each pairwise comparison (No-FS vs GA, No-FS vs PSO, No-FS vs RFE) across the Valid Range accuracy scores of all classifiers. In no comparison did the test yield a statistically significant difference (all $p > 0.05$), confirming that the accuracy differences between feature-selection conditions are not reliably distinguishable from random variation. This finding is consistent with the hypothesis that when features are broadly informative (as in the present dataset), the marginal gain from pruning uninformative variables is insufficient to produce measurable performance improvements [28]. Importantly, the inclusion of three feature-selection paradigms in the framework is not motivated by accuracy improvement claims, but by the methodological goal of systematic cross-paradigm comparison — a contribution that is itself novel in the educational data mining literature, where single-method feature selection remains the norm.

The finding that FFNN depth does not improve accuracy in this context is educationally and technically informative. It implies that the non-linear complexity of the student-performance prediction problem does not exceed what a single-layer network can approximate, given the variable types and sample size involved. This finding discourages unnecessary model complexity in similar institutional settings and provides a principled justification for preferring shallower, faster-training architectures in practice.

VI. Conclusion

This paper presented a novel seven-layer computational framework for predicting student performance ranges in higher education. The framework addresses five identified gaps in the existing literature: the absence of range-valued predictions, insufficient treatment of imbalanced data, limited application of unsupervised learning, shallow data analysis practices, and the lack of systematic neural network architecture comparisons. Validated on a rich dataset of over 200,055 graduate records from the Ibn Zohr University of Morocco, the framework achieved Valid Range accuracy of 84%/86% (worst/common case) with a 3%/5% improvement over a literature baseline of 81%. The unsupervised Popularity Architecture attained peak accuracy of 96.91%, the highest reported in this experimental context.

Five concrete contributions emerge from this work. First, the performance range output format provides a more realistic representation of student academic capability, explicitly acknowledging the influence of life circumstances on grade variability. Second, the Anchored Training Data algorithm offers a practical, computationally inexpensive method for reducing imbalanced data effects without requiring complex oversampling strategies. Third, the Popularity Architecture demonstrates that K-Modes/K-Prototype clustering can function as a direct predictive mechanism for academic performance, opening a new methodological direction in educational data mining. Fourth, the eight-FFNN comparison provides empirical evidence that increasing hidden-layer depth does not improve predictive accuracy in this context, informing future model design choices. Fifth, the comprehensive data analysis layer exposes dataset properties—including socioeconomic attainment gradients and the limitations of dimensionality reduction—that have practical implications for institutional policy.

Several directions merit future investigation. Replicating the framework on datasets from multiple institutions would strengthen external validity and enable cross-institutional comparisons. Enriching the feature space with health, financial, and detailed parental variables could expose additional predictive signals, subject to ethical and legal data-governance constraints. Extending the imbalance-mitigation layer with rule-based sampling architecture, and exploring additional clustering algorithms in the Popularity Architecture, represent tractable methodological improvements. Finally, packaging the framework as an API-accessible microservice would facilitate integration with existing student performance dashboard systems, realizing its practical impact for institutional decision-making.

References

[1] S. W. Pyke et P. Sheridan, « Logistic Regression Analysis of Graduate Student Retention », Can. J. High. Educ., vol. 23, no 2, p. 44-64, août 1993, doi: 10.47678/cjhe.v23i2.183161.

[2] A. I. Adekitan et E. Noma‐Osaghae, « Data mining approach to predicting the performance of first year student in a university using the admission requirements », Educ. Inf. Technol., vol. 24, no 2, p. 1527-1543, déc. 2018, doi: 10.1007/s10639-018-9839-7.

[3] C. F. Rodríguez-Hernández, M. Musso, E. Kyndt, et E. Cascallar, « Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation », Comput. Educ. Artif. Intell., vol. 2, p. 100018-100018, janv. 2021, doi: 10.1016/j.caeai.2021.100018.

[4] H. Waheed et al., « Balancing sequential data to predict students at-risk using adversarial networks », Comput. Electr. Eng., vol. 93, p. 107274-107274, juin 2021, doi: 10.1016/j.compeleceng.2021.107274.

[5] S. Alwarthan, N. Aslam, et I. U. Khan, « Predicting Student Academic Performance at Higher Education Using Data Mining: A Systematic Review », Appl. Comput. Intell. Soft Comput., vol. 2022, p. 1-26, sept. 2022, doi: 10.1155/2022/8924028.

[6] A. Hellas et al., « Predicting academic performance: a systematic literature review », p. 175-199, juill. 2018, doi: 10.1145/3293881.3295783.

[7] M. A. Chaudhry et E. Kazim, « Artificial Intelligence in Education (AIEd): a high-level academic and industry note 2021 », AI Ethics, vol. 2, no 1, p. 157-165, juill. 2021, doi: 10.1007/s43681-021-00074-z.

[8] V. Dignum, « The role and challenges of education for responsible AI », Lond. Rev. Educ., vol. 19, no 1, janv. 2021, doi: 10.14324/lre.19.1.01.

[9] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H. Y. Kwon, et A. Hussain, « Educational data mining to predict students' academic performance: A survey study », Educ. Inf. Technol., vol. 28, p. 905-971, janv. 2023.

[10] B. Alnasyan, M. Basheri, et M. O. Alassafi, « Deep Learning Techniques for Predicting Student's Academic Performance on Virtual Learning Environments: A Review », Research Square (Research Square). Research Square (United States), janvier 2024. doi: 10.21203/rs.3.rs-3888441/v1.

[11] S. J. H. Yang, O. H. T. Lu, A. Y. Q. Huang, J. Huang, H. Ogata, et A. J. Q. Lin, « Predicting Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis », J. Inf. Process., vol. 26, p. 170-176, janv. 2018, doi: 10.2197/ipsjjip.26.170.

[12] N. Pereira, « Factor Analysis of Mixed Data (FAMD) and Multiple Linear Regression in R », ARROWDublin Inst. Technol. Dublin Inst. Technol., janv. 2019, [En ligne]. Disponible sur: https://arrow.tudublin.ie/scschcomdis/212

[13] R. R. Utzman, D. L. Riddle, et D. V. Jewell, « Use of Demographic and Quantitative Admissions Data to Predict Performance on the National Physical Therapy Examination », Phys. Ther., vol. 87, no 9, p. 1181-1193, juill. 2007, doi: 10.2522/ptj.20060222.

[14] J. Jaber, N. Keric, P. Kang, et A. Feinstein, « Predicting success: A comparative analysis of student performance on the surgical clerkship and the NBME surgery subject exam », Surg. Open Sci., vol. 1, no 2, p. 86-89, août 2019, doi: 10.1016/j.sopen.2019.07.002.

[15] H. A. Mengash, « Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems », IEEE Access, vol. 8, p. 55462-55470, janv. 2020, doi: 10.1109/access.2020.2981905.

[16] S. D. A. Bujang, A. Selamat, et O. Krejcar, « A Predictive Analytics Model for Students Grade Prediction by Supervised Machine Learning », in IOP Conference Series Materials Science and Engineering, IOP Publishing, févr. 2021, p. 12005-12005. doi: 10.1088/1757-899x/1051/1/012005.

[17] B. J. Kiss, M. Nagy, R. Molontay, et B. Csabay, « Predicting Dropout Using High School and First-semester Academic Achievement Measures », p. 383-389, nov. 2019, doi: 10.1109/iceta48886.2019.9040158.

[18] A. Bilal, « Predicting students' academic performance based on enrolment data », in Proceedings of the 2020 Conference on Applied Machine Learning, janv. 2020, p. 54-61.

[19] N. Z. Zacharis, « Predicting Student Academic Performance in Blended Learning Using Artificial Neural Networks », Int. J. Artif. Intell. Appl., vol. 7, no 5, p. 17-29, sept. 2016, doi: 10.5121/ijaia.2016.7502.

[20] M. Ding, K. Yang, D.-Y. Yeung, et T.-C. Pong, « Effective Feature Learning with Unsupervised Learning for Improving the Predictive Models in Massive Open Online Courses », p. 135-144, févr. 2019, doi: 10.1145/3303772.3303795.

[21] S. Helal et al., « Predicting academic performance by considering student heterogeneity », Knowl.-Based Syst., vol. 161, p. 134-146, juill. 2018, doi: 10.1016/j.knosys.2018.07.042.

[22] Y. Zhao, Q. Xu, M. Chen, et G. M. Weiss, « Predicting Student Performance in a Master's Program in Data Science using Admissions Data. », in Educational Data Mining, janv. 2020. [En ligne]. Disponible sur: https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_27.pdf

[23] B. K. Francis et S. B. Suvanam, « Predicting Academic Performance of Students Using a Hybrid Data Mining Approach », J. Med. Syst., vol. 43, no 6, p. 162-162, avr. 2019, doi: 10.1007/s10916-019-1295-4.

[24] O. E. Aissaoui, Y. E. M. E. Alami, L. Oughdir, A. Dakkak, et Y. E. Allioui, « A Multiple Linear Regression-Based Approach to Predict Student Performance », in Advances in intelligent systems and computing, Springer Nature, 2020, p. 9-23. doi: 10.1007/978-3-030-36653-7_2.

[25] A. Bettahi, F.-Z. Belouadha, et H. Harroud, « A Modular and Explainable Machine Learning Pipeline for Student Dropout Prediction in Higher Education », Algorithms, vol. 18, no 10, p. 662-662, oct. 2025, doi: 10.3390/a18100662.

[26] M. Fazil, A. Rísquez, et C. Halpin, « A Novel Deep Learning Model for Student Performance Prediction Using Engagement Data », J. Learn. Anal., vol. 11, no 2, p. 23-41, mai 2024, doi: 10.18608/jla.2024.7985.

[27] B. Strøm, T. Falch, et P. Lujala, « Geographical constraints and educational attainment », RePEc Res. Pap. Econ., sept. 2011.

[28] A. Abdulla, G. Baryannis, et I. Badi, « Weighting the Key Features Affecting Supplier Selection using Machine Learning Techniques », Preprints.org, déc. 2019, doi: 10.20944/preprints201912.0154.v1.

[29] S. Greatorex-Voith et A. Anand, « A data-driven framework for identifying high school students at risk of not graduating on time », in Workshop Proceedings, EDM 2015, janv. 2015.

[30] H. Li, C. Lynch, et T. Barnes, « Early Prediction of Course Grades: Models and Feature Selection », ArXiv Cornell Univ., déc. 2018, doi: 10.48550/arxiv.1812.00843.

[31] S. Alturki, I. Hulpuș, et H. Stuckenschmidt, « Predicting Academic Outcomes: A Survey from 2007 Till 2018 », Technol. Knowl. Learn., vol. 27, no 1, p. 275-307, sept. 2020, doi: 10.1007/s10758-020-09476-0.

[32] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, et R. Nawaz, « Predicting academic performance of students from VLE big data using deep learning models », Comput. Hum. Behav., vol. 104, p. 106189-106189, nov. 2019, doi: 10.1016/j.chb.2019.106189.