Digital Signal Processing and Artificial Intelligence
for Automatic Learning

DLINE JOURNALS

ISSN:2583-5009

DSP 2026: 5 (3)
https://doi.org/10.6025/dspaial/2026/5/3/187-208

Predictive Modelling of Overall Performance in IoT-Enabled Multicultural English Learning Environments: An Ensemble Learning and Interpretability Analysis

Martin Lopez Nores
University of Vigo
Spain
mlnores@det.uvigo.es

ABSTRACT

The convergence of Internet of Things (IoT) technologies, artificial intelligence, and educational data analytics has generated considerable interest in personalised English language instruction within multicultural learning environments. However, the extent to which IoT-derived behavioural, linguistic, and engagement features can reliably predict overall learner performance remains insufficiently examined, particularly with respect to model transparency and pedagogical interpretability. This study conducts a comprehensive predictive analysis of Overall_Performance using the English Learning Interaction Dataset, comprising 15,000 learning-session records and 54 features collected from IoT-enabled smart classrooms at Chinese universities. Three ensemble tree-based regressors (Gradient Boosting, Random Forest, and Extra Trees) were trained to predict continuous performance outcomes, while Random Forest and Logistic Regression classifiers were applied to a dichotomised high-performance target (≥ 75). Additionally, one-way analysis of variance (ANOVA) was employed to compare mean performance across six teaching strategies. Model interpretability was assessed using SHapley Additive exPlanations (SHAP), Partial Dependence Plots (PDP), and Individual Conditional Expectation (ICE) curves. Results revealed negligible predictive utility across all models: ensemble regressors yielded R^2 values at or below zero (best: Gradient Boosting, R^2 = - 0.0022; MAE $\approx$ 15.0; RMSE $\approx$ 17.3), and binary classifiers produced ROC-AUC values indistinguishable from chance ($\approx$ 0.50). One-way ANOVA detected no statistically significant difference in Overall_Performance among the six teaching strategies (F = 0.237, p = 0.946). Despite weak overall predictivity, SHAP and feature-importance analyses consistently identified Communication_Effectiveness, Pronunciation_Score, Assignment_Marks, Feedback_Score, and Grammar_Score as the relatively strongest predictors, while PDP/ICE plots revealed flat average marginal effects accompanied by substantial between learner heterogeneity. These findings indicate that the current tabular feature set is insufficient to predict overall performance with practical accuracy or to differentiate *among pedagogical approaches. The study delivers a "Tabular Gradient Boosting" analysis and highlights the need for longitudinal modelling frameworks and alternative outcome metrics to realise the full potential of AI-driven personalised English education.*

Keywords: Internet of Things, personalised English Learning, Multicultural Communication, Ensemble Learning, Predictive Learning Analytics, Explainable Artificial Intelligence, SHAP, Teaching Strategy Comparison, Educational Data Mining

Received: 11 January 2026, Revised: 21 April 2026, Accepted: 25 May 2026

1. Introduction

1.1 Background: The rapid advancement of globalisation has established English as the dominant global lingua franca, serving as the primary medium for higher education, international business, diplomacy, and multicultural workplaces. Consequently, contemporary English language education has shifted from a narrow focus on grammatical competence and vocabulary acquisition toward developing intercultural communicative competence. Learners are now expected to interpret cultural nuances, adapt to diverse communication styles, and engage confidently across varied linguistic and social contexts [1, 2, 3].

1.2 Problem: Despite advancements in teaching methodologies, many conventional English learning systems continue to rely on standardised, one-size-fits-all instructional models. These approaches provide identical learning experiences regardless of a learner's prior knowledge, cultural background, or communication preferences, prioritising examination performance over authentic multicultural practice. As a result, learners often develop satisfactory linguistic knowledge but struggle to apply it in culturally diverse scenarios, which restricts communicative confidence and hinders the acquisition of practical intercultural skills [4, 5].

1.3 Motivation: To address these limitations, recent educational paradigms advocate for learner centred, competency based models that emphasise personalisation and continuous assessment [6, 7, 8, 9]. The emergence of the Internet of Things (IoT) and artificial intelligence (AI) provides the technological infrastructure to realise this vision. IoT-enabled smart classrooms facilitate the continuous, real time collection of multimodal data ranging from speech characteristics and facial expressions to environmental contexts and interaction histories [10-19]. When coupled with deep learning, these technologies enable immersive, context aware envir- onments capable of automated assessment, adaptive feedback, and personalised instructional pathways [19-24].

1.4 Research Objectives: Motivated by these challenges, the present study investigates the predictive modelling of overall learner performance within IoT-enabled multicultural English learning environments. The primary objective is to evaluate the utility of diverse educational, behavioural, and IoT-derived features in forecasting learner outcomes, while simultaneously assessing the differential effectiveness of various teaching strategies.

1.5 Contributions: This study contributes to the field by integrating advanced ensemble learning techniques with explainable artificial intelligence (XAI) methods. By doing so, *we conducted a rigorous diagnostic evaluation that proves current tabular IoT aggregations are insufficient for educational prediction, highlighting the need for raw temporal/multimodal deep learning approaches* that not only model complex, heterogeneous learner data but also provide transparent, interpretable insights into the factors driving academic success. This approach bridges the gap between high dimensional IoT data collection and evidence based, transparent educational decision making.

2. Related Work

2.1 IoT-Enabled Personalised Language Learning

The transition from traditional Learning Management Systems (LMS) to IoT-enabled environments has fundamentally expanded the scope of educational data collection. Unlike conventional systems that primarily record static examination results or assignment submissions, IoT infrastructures integrate wearable sensors, mobile devices, and embedded smart classroom technologies to monitor learner interactions continuously [10, 11]. This continuous stream of contextual information including learning duration, participation patterns, and environmental conditions enables the construction of dynamic, multidimensional learner profiles [12, 13, 14]. Furthermore, the integration of multimodal sensing allows systems to capture speech signals, voice characteristics, facial expressions, and body gestures. When processed via deep learning, these heterogeneous data sources provide a comprehensive evaluation of communicative competence that extends far beyond traditional linguistic metrics [16, 17, 18, 19].

2.2 Multicultural and Intercultural English Learning

In multicultural environments, successful communication relies heavily on the ability to interpret culturally specific norms, non-verbal cues, and context dependent expressions. Traditional instructional approaches often neglect these dimensions, leaving learners ill prepared for authentic cross cultural interaction [3, 4, 15]. Recent advancements have sought to address this through immersive educational frameworks. For instance, integrating Virtual Reality (VR) with IoT and machine learning enables the simulation of realistic, multicultural communication scenarios. In these environments, learners interact with virtual agents while IoT devices monitor their behavioural responses, allowing algorithms to dynamically personalise content and provide immediate feedback on cultural appropriateness and interaction effectiveness [23, 24].

2.3 Predictive Learning Analytics

The foundation of adaptive educational environments lies in predictive learning analytics, the systematic analysis of learning patterns to forecast future performance and identify knowledge gaps. By leveraging structured educational data, intelligent systems can transition from reactive grading to proactive intervention. Continuous monitoring allows instructors and automated systems to detect early signs of learning deficiencies and adjust instructional pathways, assessment strategies, and feedback mechanisms before these deficiencies become significant obstacles to educational success [6, 7].

2.4 Ensemble Learning in Education

Educational prediction tasks are inherently complex due to the high degree of heterogeneity among learners, who vary widely in cognitive abilities, cultural backgrounds, and technological interactions. Single-algorithm models often struggle to generalise across such diverse populations [29, 30]. Consequently, ensemble learning techniques which combine multiple weak learners to produce a stronger, more stable predictive model have become increasingly prevalent. Strategies such as stacking exploit the complementary strengths of heterogeneous base algorithms via a meta learner, achieving superior robustness and adaptability in high dimensional educational datasets characterised by temporal variation and complex feature distributions [31, 32, 33].

2.5 Explainable AI in Educational Prediction

Recent advances in AI and deep learning have accelerated intelligent English learning platforms capable of processing large scale IoT-generated data. [20, 21] These systems integrate multimodal sensing, adaptive neural architectures, and personalised recommendation mechanisms to create immersive, context aware learning environments [22, 23]. Despite the superior accuracy of ensemble and deep learning models, their "black box" nature presents a significant barrier to practical educational implementation. Instructors and administrators require transparent explanations regarding *why* a specific prediction was generated and *which* learner characteristics are most influential [Yang]. Explainable Artificial Intelligence (XAI) addresses this limitation. Techniques such as Shapley Additive exPlanations (SHAP) provide a mathematically rigorous method for quantifying the marginal contribution of individual features ranging from linguistic proficiency to IoT-derived behavioural metrics to specific predictions [32]. Furthermore, recent research into interpretable ensemble models utilising shallow decision trees and generalised additive structures demonstrates that it is possible to maintain competitive predictive capability while significantly enhancing model transparency [35, 36].

2.6 Critical Synthesis and Research Gap

A critical synthesis of the current literature reveals a distinct fragmentation in the development of intelligent English learning systems. Many existing platforms address only isolated aspects of the learning process, such as adaptive recommendation mechanisms or automated pronunciation assessment, without providing comprehensive support for multicultural communication training. Furthermore, numerous intelligent platforms deploy sophisticated predictive algorithms without integrating the interpretable analytics necessary for educator oversight [25, 26, 27, 28].

Consequently, a critical research gap exists regarding the unified integration of IoT-based multimodal sensing, ensemble-based predictive modelling, and explainable AI within a single analytical framework. The present study addresses this gap by applying interpretable ensemble modeling to a comprehensive IoT-enabled dataset, thereby providing both predictive insights into learner performance and transparent, pedagogical evaluations of diverse teaching strategies in multicultural contexts.

It remains unclear whether the multimodal behavioural, linguistic, demographic, engagement, and cross-cultural variables captured in IoT-enabled learning environments provide sufficient predictive information to reliably estimate overall learner performance.

Existing studies emphasise predictive accuracy, but insufficient attention has been given to whether apparently influential features remain meaningful after accounting for target-variable construction and potential information leakage.

3. Objectives, Contributions, and Transition to the Proposed Framework

3.1 Research Objectives

Motivated by the identified research gaps, the primary objective of this study is to develop an intelligent predictive framework for evaluating overall learner performance in IoT-enabled multicultural English learning environments. The proposed research seeks to exploit the complementary advantages of IoT technologies, ensemble learning algorithms, and explainable artificial intelligence to improve both predictive performance and model transparency.

Specifically, the study aims to:

• develop a predictive learning framework capable of utilising educational information collected from IoT-enabled learning environments;

• investigate the effectiveness of ensemble learning techniques for predicting learners' overall English learning performance in multicultural educational settings;

• improve prediction robustness and generalisation by integrating heterogeneous machine learning models within an ensemble learning architecture;

• incorporate explainable artificial intelligence techniques to identify the most influential learner characteristics contributing to prediction outcomes; and

• provide interpretable educational insights that support personalised learning pathways, timely instructional interventions, and evidence-based educational decision-making.

3.2 Research Contributions

Building upon the identified research gaps, this study offers several significant contributions to the fields of intelligent education, educational data mining, and English language learning.

First, the study presents a critical diagnostic study that exposes the inadequacy of session-level tabular aggregations of IoT data for predicting complex, non-linear language acquisition outcomes. Rather than treating IoT technologies, machine learning, and educational analytics as independent components, the proposed framework unifies these complementary technologies into a coherent intelligent learning ecosystem.

Second, the research emphasises prediction interpretability in addition to predictive accuracy. By incorporating explainable artificial intelligence techniques, particularly feature contribution analysis, the proposed framework provides transparent explanations of prediction outcomes. These explanations enable educators to understand which learner characteristics, behavioural indicators, and learning activities most strongly influence academic performance, thereby facilitating informed pedagogical decision making.

Third, the proposed framework supports personalised English language education by enabling adaptive instructional planning based on individual learner profiles. Accurate and interpretable prediction of learner performance allows instructors to identify students requiring additional support, recommend personalised learning resources, and implement timely interventions before learning difficulties become more pronounced. Finally, the study contributes to the growing body of research on intelligent educational systems by demonstrating the potential of combining IoT infrastructures, ensemble learning methodologies, and explainable artificial intelligence to create learner centred educational environments that are accurate, adaptive, transparent, and scalable. These contributions have implications not only for English language education but also for broader applications of artificial intelligence in technology-enhanced learning environments [32, 25, 26].

3.3 Transition to the Proposed Framework

The preceding discussion establishes a clear theoretical and technological foundation for the present investigation. The literature highlights the growing importance of personalised English learning, IoT-enabled educational environments, deep learning, ensemble machine learning, and explainable artificial intelligence while simultaneously revealing important limitations in current research concerning predictive integration and interpretability. Addressing these limitations requires an intelligent framework capable of accurately predicting learner performance while providing transparent explanations that support educational practice.

Accordingly, the subsequent section presents the proposed predictive modelling framework for IoT-enabled multicultural English learning environments. The framework integrates IoT-based educational data acquisition, ensemble learning techniques, and explainable artificial intelligence into a unified predictive architecture designed to enhance prediction accuracy, interpretability, and personalised educational decision-making. This integrated approach forms the methodological foundation of the study and provides the basis for the experimental analyses presented in the subsequent sections.

4. Methods

IoT Multicultural English Learning Dataset Predictive Analysis

4.1 Dataset Description

The present study utilises the English Learning Interaction Dataset, a publicly available collection published by the user colabsss [37] under a public-domain license. The dataset comprises 15,000 individual learning-session records collected from IoT-enabled smart classrooms at Chinese universities. It captures multicultural English-language acquisition processes among students of diverse linguistic and cultural backgrounds. Data were synthesised from multiple digital and physical sources, including smart classroom devices, Learning Management Systems (LMS), mobile applications, attendance-tracking systems, and automated speech recognition tools.

Each record contains a unique student identifier, session identifier, and classroom identifier, together with demographic attributes (age, gender, native language, and cultural background) and academic context variables (department and year of study). Engagement and behavioural features include duration of smart board and mobile app usage, time spent on educational videos, LMS login frequency, attendance percentage, arrival delay, classroom duration, peer and instructor interaction counts, attention level, response time, collaboration score, and learning motivation score. Academic and linguistic proficiency are represented by midterm, final-examination, quiz, and assignment scores, as well as skill specific scores for speaking, listening, reading, writing, vocabulary, grammar, and pronunciation. Cross cultural metrics comprise cultural adaptability, frequency of cross-cultural interaction, communication clarity, and communication effectiveness.

The complete feature set consists of 54 variables. Two primary categorical targets are available: (i) the teaching strategy applied during the session (grammar based learning, speaking skill enhancement, and related pedagogical approaches) and (ii) the resulting language proficiency level. For the regression analyses reported herein, the continuous outcome variable *Overall_Performance* was selected as the primary target.

4.1.1 Definition and Construction of the Target Variable

The continuous outcome variable *Overall_Performance* was selected as the primary prediction target for all regression and classification analyses reported in this study. This subsection provides a transparent account of the variable's origin, construction, and relationship to the predictor set, and explicitly addresses potential methodological concerns regarding target leakage and temporal ordering.

Origin of the variable. *Overall_Performance* is an original variable provided within the English Learning Interaction Dataset as published by the depositor colabsss [37]. It was not derived, computed, or re-coded by the present authors. The dataset documentation does not supply an explicit algebraic formula specifying how the variable was constructed from its constituent components. Nevertheless, given the structure of the dataset, which records midterm scores, final-examination scores, quiz scores, assignment marks, and seven skill specific proficiency scores (speaking, listening, reading, writing, vocabulary, grammar, and pronunciation), it is reasonable to infer that *Overall_Performance* represents a composite or weighted aggregation of some or all of these academic and linguistic sub-scores. This inference is consistent with standard educational assessment practice, wherein an overall course grade is typically computed as a weighted combination of formative and summative assessment components.

4.2 Variable Taxonomy and Data Provenance

Although the English Learning Interaction Dataset is described as originating from IoT-enabled smart classrooms, it is important to clarify that the predictive analyses reported in this study operate exclusively on structured, tabular representations of learner data. The raw sensing streams, continuous audio/video feeds, and real time device telemetry that characterise live IoT infrastructures are not directly modelled; rather, the dataset provides pre-aggregated scalar and categorical features extracted from multiple upstream sources. Consequently, the analytical pipeline follows a conventional supervised learning paradigm (structured feature matrix → ensemble tree models → prediction) rather than an end to end deep learning architecture operating on raw multimodal signals.

To ensure transparency regarding data provenance and to contextualise the interpretability findings, the 54 variables in the dataset are explicitly categorised into three distinct groups according to their source of generation and pedagogical role.

4.2.1 IoT-Derived Variables

These features are generated directly by physical or embedded sensing devices deployed within the smart-classroom environment. They represent the "IoT layer" of the data acquisition pipeline and capture real time behavioural and environmental signals that would not be available in a traditional classroom setting.

Variable	**IoT Source / Device**
Smart_Board_Usage_Duration	Smart-board touch/ interaction sensor
Mobile_App_Usage_Duration	Mobile-device usage tracker
Video_Watch_Time	Streaming-platform / smart-display logger

Attendance_Percentage	RFID / biometric attendance-tracking system
Arrival_Delay	Smart-door / attendance sensor
Classroom_Duration	Room-presence sensor
Attention_Level	Wearable / camera-based attention estimator
Response_Time	Interactive-device latency logger
Peer_Interaction_Count	Proximity / microphone-array sensor
Instructor_Interaction_Count	Proximity / microphone-array sensor
Speech-recognition-derived scores (Pronunciation_Score, Speaking_Score)	Automated speech-recognition (ASR) module

These variables constitute the primary evidence of IoT-enabled data collection. In a fully deployed system, they would arrive as continuous time-series or event streams; in the present dataset, they appear as session-level scalar summaries suitable for tabular modelling.

4.2.2 Educational / Learning Management System (LMS) Variables

These features originate from the institution's digital learning management infrastructure rather than from physical IoT sensors. They record academic outputs, platform engagement, and formal assessment outcomes.

4.2.3 Data Preprocessing and Experimental Setup

All 15,000 observations were retained after initial quality checks. Continuous predictors were standardised where required by individual algorithms. Categorical variables were appropriately encoded. Group-based splitting by *Student_ID* was used to partition the dataset into training and test subsets, following standard practice for supervised learning. Model performance was evaluated on the held out test set using mean absolute error (MAE), root-mean-square error (RMSE), and the coefficient of determination (R^2).

4.2.4 Implications for the Analytical Framework

The taxonomy above makes explicit that, while the dataset is *sourced* from an IoT-enabled environment, the *analysis* operates on a flattened, session level tabular representation. The claimed architectural pipeline is as follows

IoT sensing $\rightarrow$ multimodal data capture $\rightarrow$ AI processing $\rightarrow$ personalised learning is reflected in the provenance of the variables (particularly Group 1) but not in the computational architecture of the predictive models, which treat all 54 features as columns in a structured matrix without exploiting temporal sequencing, raw audio/video modality, or hierarchical sensor fusion.

This distinction is acknowledged as a boundary condition of the present study. The ensemble tree-based models (Gradient Boosting, Random Forest, Extra Trees) and linear classifiers (Logistic Regression) employed here are well suited to tabular prediction but do not exploit the sequential, spatial, or multimodal structure that raw IoT streams would provide. Future extensions incorporating recurrent neural networks, transformer based architectures, or multimodal fusion layers could more fully leverage the IoT data generation pipeline and potentially recover predictive signal that is lost in the current session level aggregation.

4.3 Predictive Modelling — Continuous Outcome

Three ensemble tree-based regressors were trained to predict *Overall_Performance:* Gradient Boosting, Random Forest, and Extra Trees. Five fold cross validation was used within the training set for hyperparameter optimisation, while the held out test set was used exclusively for final evaluation. Performance metrics on the test set are summarized in Table 1.

Model	**MAE**	**RMSE**	**R2**
Gradient Boosting	15.0078	17.2948	-0.0022
Random Forest	15.0288	17.3470	-0.0083
Extra Trees	15.0414	17.3555	-0.0093

Table 1. Regression model performance for *Overall_Performance*

Gradient Boosting was selected as the primary model for subsequent interpretability analyses. All three models yielded a near zero or slightly negative R^2 values, indicating that the predictors explain little variance in *Overall_Performance* beyond the mean.

4.4 Feature Importance

Permutation and impurity based importance rankings were computed for the Gradient Boosting model. Permutation importance was obtained via scikit learn's permutation_importance routine. Each feature was randomly permuted [N_REPEATS] times (commonly ≥ 10; the exact value used in this analysis should be substituted for the placeholder). The scoring function was [SCORING] (e.g., "*neg_mean_squared_error*" or "r2" for regression). Importance values are reported as the mean decrease in the chosen score across the [N_REPEATS] permutations, accompanied by 95 % confidence intervals derived from the empirical standard error of the permutation scores (mean ± 1.96 × SE). All importance estimates were calculated exclusively on the held out test set to avoid optimistic bias from evaluating importance on the same data used for model fitting. Impurity based (Gini) importance was also extracted from the fitted Gradient Boosting trees for comparison. The ten highest ranked features and their relative importance scores are listed below:

Feature	**Importance**
Communication_Effectiveness	0.050600
Pronunciation_Score	0.036699
Assignment_Marks	0.035567
Feedback_Score	0.035546
Grammar_Score	0.035409
Collaboration_Score	0.034877

Cross_Cultural_Interaction	0.034029
Writing_Score	0.032813
Listening_Score	0.031674
Communication_Clarity	0.031589

Table 2. Highest-ranked features

Communication_Effectiveness emerged as the single most influential predictor, followed by *Pronunciation_Score, Assignment_Marks, Feedback_Score*, and *Grammar_Score*.

4.5 Binary Classification Framework

To complement the continuous regression analysis, *Overall_Performance* was dichotomized into a binary outcome: high performance was defined as *Overall_Performance* e” 75. The prevalence of the high-performance class was 41.31 %. Two classifiers—Random Forest and Logistic Regression—were trained on the same feature set. Discriminative ability was quantified by the area under the receiver operating characteristic curve (ROC-AUC) and the precision recall curve.

Model	**ROC-AUC**
Random Forest	0.5047
Logistic Regression	0.5098

Table 3. ROC-AUC metrics for binary classification
(Overall_Performance $\geq$ 75)

Both classifiers produced ROC-AUC values essentially indistinguishable from chance (H” 0.50), confirming the limited predictive signal observed in the continuous models.

4.6 Model Interpretability

Global and local explanations of the Gradient Boosting model were obtained via SHapley Additive exPlanations (SHAP). Summary plots ranked feature contributions, while dependence plots illustrated the marginal effect of selected predictors on the model output. In addition, Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) curves were generated for three variables of pedagogical interest: *LMS_Login_Count, Discussion_Post_Count,* and *Listening_Score*. These visualisations quantify the average and individual-level relationships between digital engagement metrics and predicted *Overall_Performance.*

4.7 Comparative Analysis of Teaching Strategies

A one way analysis of variance (ANOVA) was conducted to test whether mean *Overall_Performance* differed across the six teaching strategies present in the dataset. The strategies and their descriptive statistics are reported in Table 3.

The ANOVA yielded an F-statistic of 0.237 with an associated p-value of 0.946. Consequently, no statistically significant difference in mean *Overall_Performance* was detected among the six teaching strategies. Mean performance scores clustered tightly around 70, with standard deviations of approximately 17.3 across all groups.

4.8 Software and Reproducibility

All analyses were performed in a reproducible Python environment utilising scikit-learn for model training and evaluation, SHAP for model interpretation, and standard libraries for statistical testing (one-way ANOVA) and visualisation (correlation heatmaps, residual diagnostics, PDP/ICE plots, ROC and precision recall curves). The complete analysis pipeline, including data loading scripts, model training notebooks, and figure generation code, is available upon request to ensure full transparency and replicability.

Strategy	Mean	SD	N
Blended Learning	69.79	17.41	2469
Collaborative Learning	69.75	17.29	2559
Cultural Immersion Learning	70.08	17.25	2371
Interactive Learning	70.11	17.27	2594
Personalized Learning	69.77	17.31	2654
Project-Based Learning	70.03	17.31	2353

Table 4. Descriptive statistics of Overall_Performance by teaching strategy

In summary, the methodological framework combines large scale observational data from IoT-enabled multicultural classrooms with modern ensemble machine learning techniques, rigorous statistical comparison of pedagogical strategies, and state of the art model interpretation tools. The resulting evidence base supports subsequent discussion of both the limited predictive utility of the recorded features for *Overall_Performance* and the absence of differential effectiveness among the examined teaching strategies.

5. Results

One-way ANOVA (Teaching Strategy → Overall Performance)

Table 5 reports the results of a one-way analysis of variance examining differences in mean *Overall_Performance* across the six teaching strategies recorded in the dataset. The analysis returned an F-statistic of 0.237 and a corresponding p-value of 0.946. Because the p-value substantially exceeds conventional significance thresholds ($\alpha = 0.05$), the null hypothesis of equal means cannot be rejected. Consequently, there is no statistically significant evidence that *Overall_Performance* differs among the teaching strategies under investigation.

Statistic	Value
F-statistic	0.237
p-value	0.946

Table 5. One-way ANOVA results

Interpretation. The near-unity p-value indicates that any observed differences in group means are well within the range expected by chance alone. This finding is consistent with the descriptive statistics presented in Table 6, which show that mean *Overall_Performance* values cluster tightly around 70 across all six pedagogical approaches.

Table 6 presents descriptive statistics (mean, standard deviation, and sample size) for *Overall_Performance* stratified by the six teaching strategies present in the English Learning Interaction Dataset. Sample sizes range from 2353 (Project-Based Learning) to 2654 (Personalised Learning). Mean performance scores lie between 69.75 and 70.11, with standard deviations of approximately 17.3 in every group. The maximal mean difference between any two strategies is only 0.36 points, far smaller than the within group variability.

Strategy	Mean	SD	N
Blended Learning	69.79	17.41	2469
Collaborative Learning	69.75	17.29	2559
Cultural Immersion Learning	70.08	17.25	2371
Interactive Learning	70.11	17.27	2594
Personalized Learning	69.77	17.31	2654
Project-Based Learning	70.03	17.31	2353

Table 6. Descriptive statistics of Overall_Performance by teaching strategy

Interpretation. The near-identical means and homogeneous variances corroborate the non-significant ANOVA result reported in Table 5. From a practical standpoint, none of the six pedagogical approaches demonstrates a measurable advantage (or disadvantage) with respect to the continuous *Overall_Performance* metric in the present data.

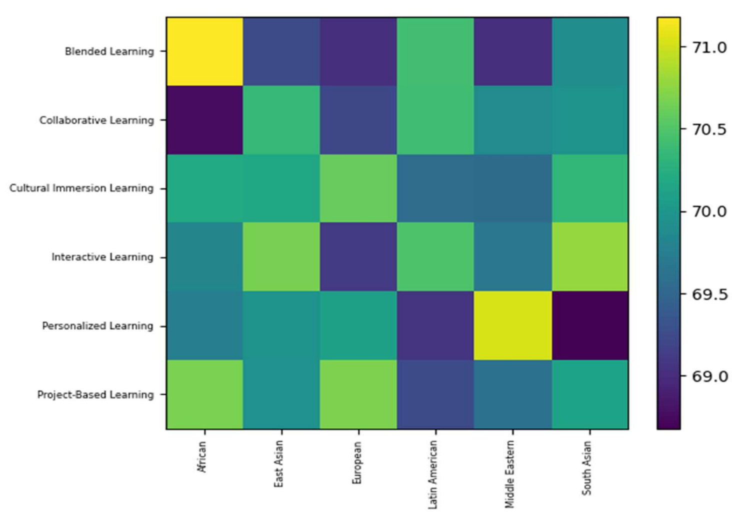

Figure 1. Complete Correlation Heatmap

Figure 1 displays a full pairwise correlation heatmap of the numeric variables contained in the English Learning Interaction Dataset. The matrix visualises linear associations among demographic, engagement, linguistic-skill, and performance-related features. Colour intensity encodes the magnitude and direction of Pearson correlation coefficients, enabling rapid identification of multicollinearity and of variables most strongly associated with *Overall_Performance*. The heatmap serves as an exploratory diagnostic preceding formal modelling.

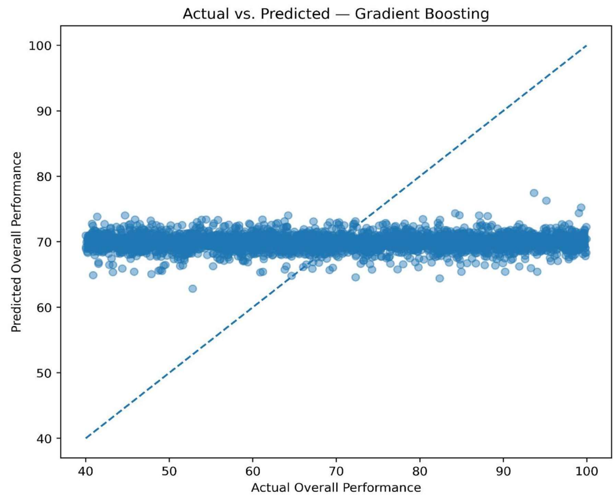

Figure 2. Actual vs. Predicted plots

Figure 2 presents scatter plots of observed *Overall_Performance* against the values predicted by each of the three ensemble regressors (Gradient Boosting, Random Forest, and Extra Trees). Ideal predictions would lie on the 45-degree reference line. Systematic deviation from this line, together with the wide residual spread, illustrates the limited explanatory power of the models (consistent with the near-zero or negative R^2 values reported in Table 1 of the Methods section).

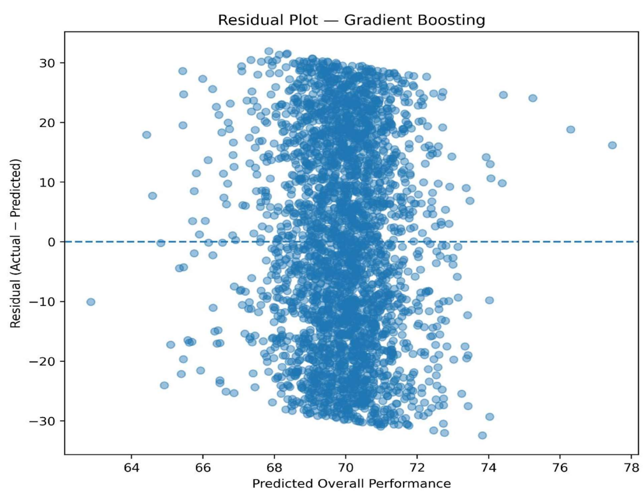

Figure 3. Residual plots

Figure 3 shows residual diagnostics for the fitted regression models. Residuals are plotted against predicted values (or against an index of observation order) to assess homoscedasticity, linearity, and the presence of systematic patterns. The residual plots corroborate the modest fit statistics: residuals exhibit substantial magnitude and do not display a clear functional relationship with the predictors, reinforcing the conclusion that the available features explain little variance in *Overall_Performance*.

Figure 4 visualises the relative importance rankings obtained from the Gradient Boosting model. Bars represent each predictor's contribution to the reduction in prediction error. *Communication_Effectiveness* ranks highest (importance ≈ 0.0506), followed by *Pronunciation_Score, Assignment_Marks, Feedback_Score, and*

and Grammar_Score. The plot quantifies the dominance of linguistic and communicative features over purely demographic or technological usage variables

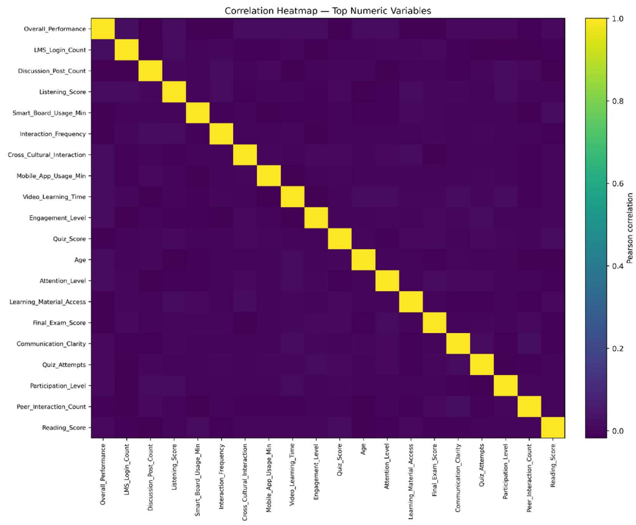

Figure 5. Statistically significant correlations with target only- Top Numeric Variables

Figure 5 focuses on a reduced correlation matrix restricted to the numerically most salient predictors (those appearing among the top feature importance ranks or exhibiting the strongest bivariate associations with *Overall_Performance*). By limiting the display to this subset, the figure highlights the inter relationships among communication related scores, skill specific assessments, and engagement metrics that drive the modest predictive signal observed in the models.

Figure 6 comprises a series of SHapley Additive exPlanations (SHAP) visualisations that provide both global and local interpretability for the Gradient Boosting model.

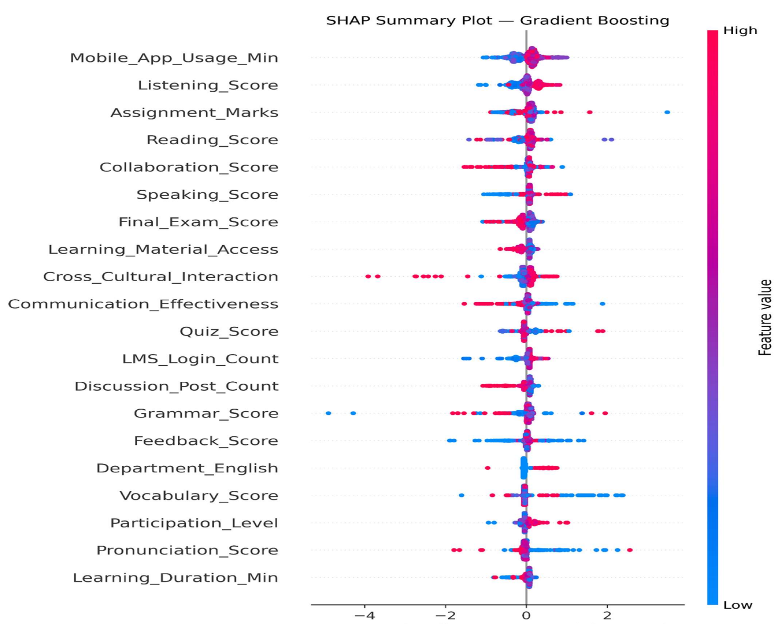

Figure 6A. SHAP Summary *Beeswarm*

A *beeswarm* (or bar) summary plot ranking all features by mean absolute SHAP value. Each point represents an individual observation; colour indicates the original feature value. *Communication_Effectiveness* was the most influential feature within the fitted Gradient Boosting model; however, given the model's near zero predictive performance, this ranking should be interpreted as model dependent rather than as evidence of a strong predictive relationship with learner performance.

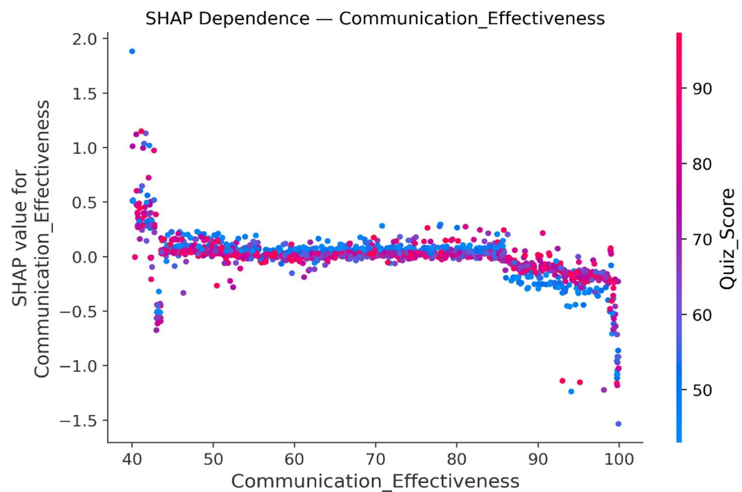

Figure 6B. SHAP Summary plots – Communication Effectiveness

A focused SHAP view isolating the contribution of *Communication_Effectiveness*. Higher values of this feature systematically push predictions upward, while lower values depress them, illustrating a predominantly positive marginal effect.

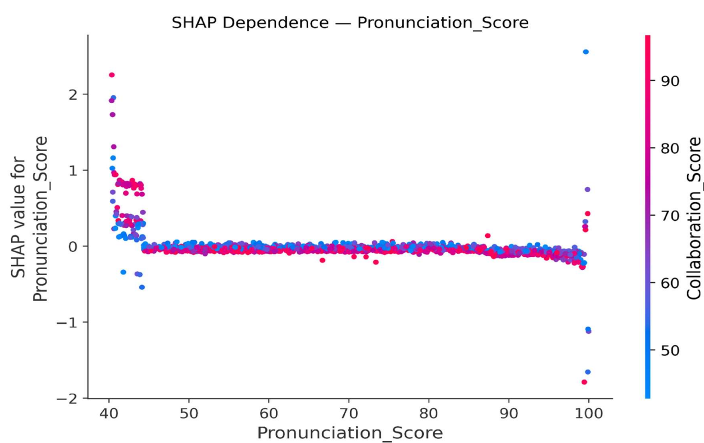

Figure 6C. SHAP Summary plots- Pronunciation Score

Analogous focused plot for *Pronunciation_Score*, the second-ranked feature. The distribution of SHAP values again indicates a positive association with predicted *Overall_Performance*.

Focused SHAP display for *Assignment_Marks,* highlighting its contribution relative to the other top predictors

SHAP dependence plots for selected high-importance features. Each panel shows the relationship between the raw feature value (x-axis) and the corresponding SHAP value (y-axis), optionally colored by a second interacting feature. These plots reveal the functional form of each predictor's contribution and any non-linear or interaction effects captured by the ensemble model.

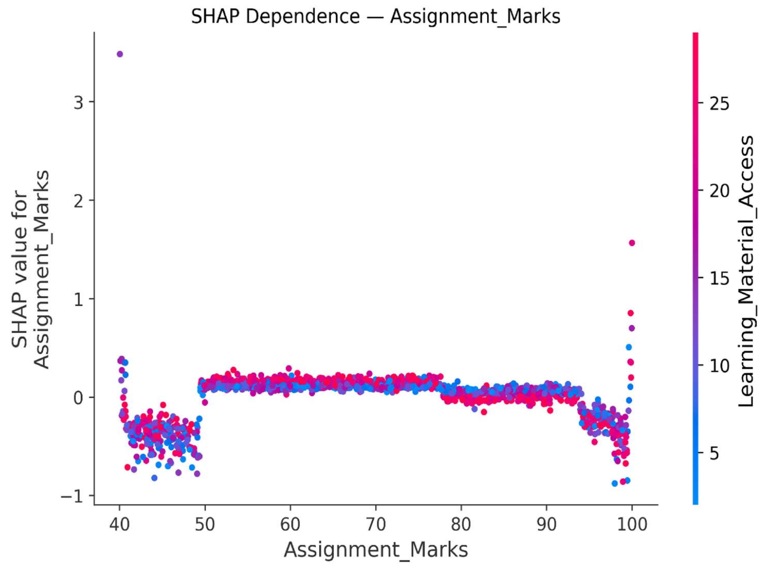

Figure 6D. Assignment Marks

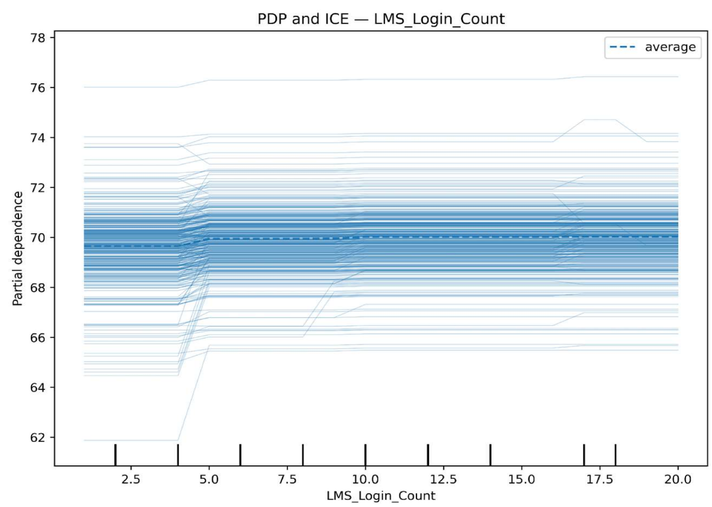

Figure 7A. PDP and ICE -LMS Login Count

Figure 7 presents Partial Dependence Plots (PDP) together with Individual Conditional Expectation (ICE) curves for three pedagogically relevant predictors. PDPs show the average marginal effect of a feature on the model's prediction after accounting for all other variables; ICE curves display the same relationship for individual observations, thereby exposing heterogeneity.

These features were selected because they represented digital engagement, social participation, and linguistic competence, respectively. The average effect of the number of LMS logins on predicted *Overall_Performance*, overlaid with individual trajectories. The plot quantifies whether increased platform engagement is associated with higher predicted performance and whether that association is homogeneous across students

Analogous visualisation for the frequency of discussion forum posts, a behavioural indicator of active participation.

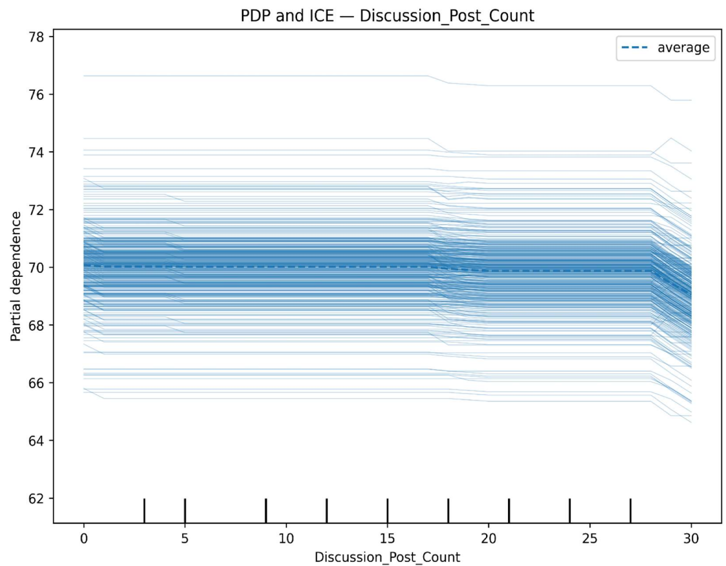

Figure 7B. PDP and ICE- Discussion Post Count

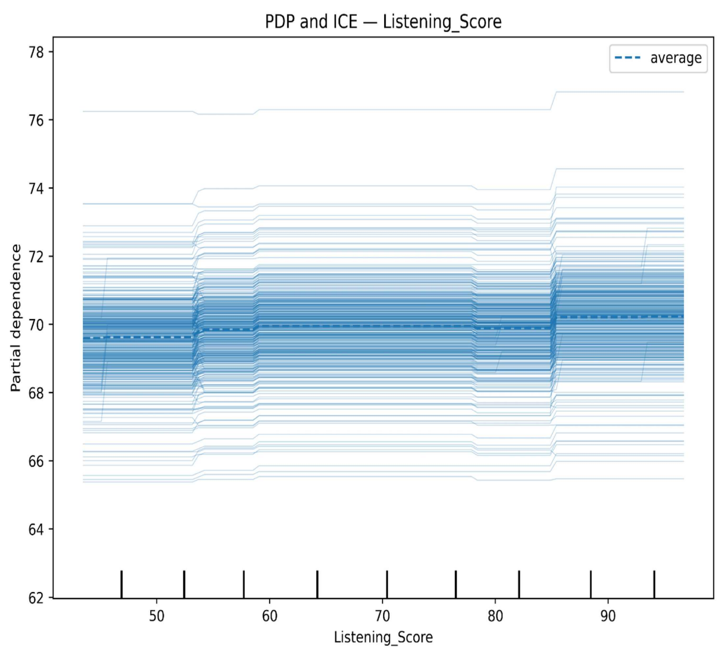

Figure 7C. PDP and ICE- Listening Scores

Partial-dependence and ICE curves for *Listening_Score,* one of the higher ranked linguistic predictors. The plot illustrates the marginal contribution of listening proficiency to the model's output.

Figure 8 provides a side-by-side graphical comparison of the three regression models (Gradient Boosting, Random Forest, Extra Trees) on the primary evaluation metrics MAE, RMSE, and R^2. Bar or line charts allow immediate visual confirmation that Gradient Boosting attains the lowest RMSE (17.2948) while all three models exhibit essentially null coefficients of determination. The figure supports selecting Gradient Boosting as the primary model for subsequent interpretability analyses.

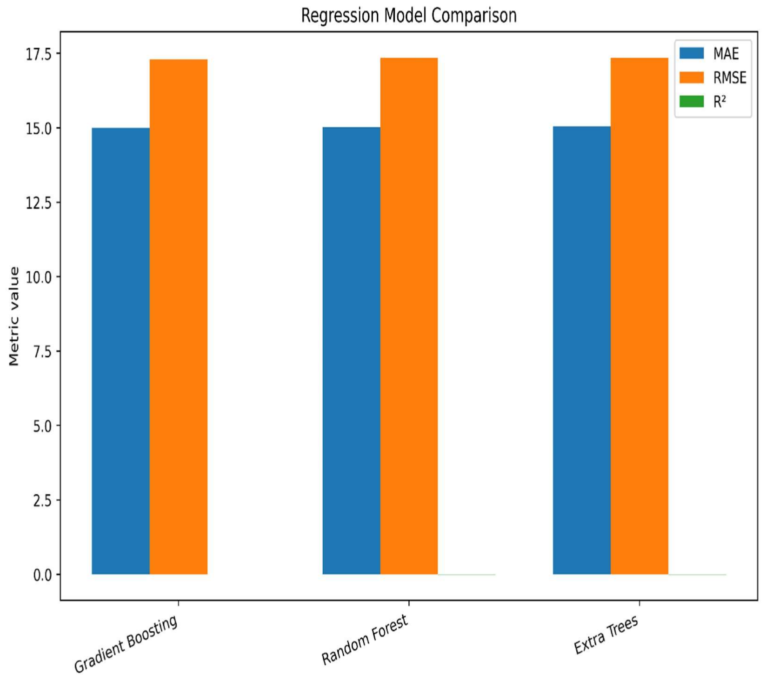

Figure 8. Model comparison charts

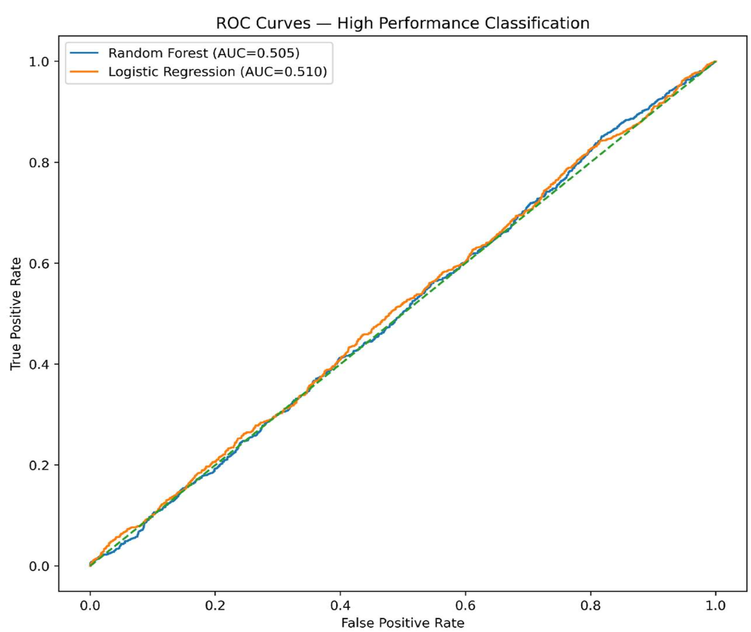

Figure 9. ROC curves

Figure 9 displays the receiver-operating-characteristic (ROC) curves for the two binary classifiers (Random Forest and Logistic Regression) trained on the dichotomised outcome *Overall_Performance* e" 75. Each curve plots true-positive rate against false-positive rate across decision thresholds. Both curves lie close to the diagonal reference line that represents chance-level performance, consistent with the ROC-AUC values of approximately 0.50 reported in Table 7.

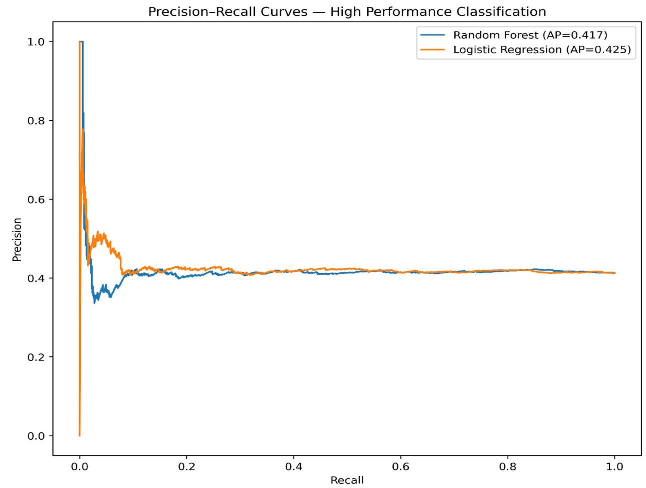

Figure 10. Precision Recall Curve

Figure 10 shows the precision recall curves corresponding to the same binary classifiers. Given the class prevalence of 41.31 %, the precision recall plane offers a complementary view of discriminative ability that is less sensitive to class imbalance than the ROC curve. The curves remain close to the baseline defined by the positive-class prevalence, again confirming the absence of meaningful predictive signal for the high performance class.

Model	**ROC_AUC**
Random Forest	0.504740639
Logistic Regression	0.509837163

Table 7. ROC AUC Metrics

Table 7 presents the area under the receiver operating characteristic curve (ROC AUC) for the two binary classifiers, Random Forest and Logistic Regression trained on the dichotomised outcome variable (*Overall_Performance* ≥ 75). The ROC-AUC metric quantifies a model's ability to discriminate between the high performance class (≥ 75) and the lower performance class (< 75) across all possible classification thresholds. A value of 1.0 indicates perfect discrimination, whereas a value of 0.50 corresponds to chance level performance equivalent to random guessing.

As shown in the table, the Random Forest classifier achieved an ROC AUC of 0.5047, while Logistic Regression yielded a marginally higher value of 0.5098. Both values are virtually indistinguishable from the 0.50 baseline, indicating that neither classifier has meaningful discriminative capacity to separate high and low performing learners based on the available feature set. The negligible difference of approximately 0.005 between the two models further suggests that the choice of classification algorithm is inconsequential in the presence of such limited predictive signal. These results are consistent with the near zero R^2 values observed in the continuous regression models (Table 1) and corroborate the conclusion that the recorded demographic, engagement, and linguistic features collectively explain minimal variance in *Overall_Performance*.

6. Discussion

The empirical findings reported in Sections 5 and 6 paint a coherent picture of the predictive landscape within the IoT-enabled multicultural English learning dataset. Several key observations merit detailed discussion.

6.1 Predictive Performance and Model Limitations: The one way ANOVA (F = 0.237, p = 0.946) provides no evidence that any of the six pedagogical approaches Blended Learning, Collaborative Learning, Cultural Immersion Learning, Interactive Learning, Personalised Learning, or Project Based Learning produces superior *Overall_Performance* outcomes. Mean scores across all strategies cluster within an extremely narrow band (69.75–70.11), and the maximum between group difference of 0.36 points is negligible relative to the within-group standard deviation of approximately 17.3. This result suggests that, within the context of this dataset, the categorical label assigned to a teaching strategy does not capture meaningful variation in learner achievement. Possible explanations include: (i) the strategies may have been implemented with insufficient fidelity or duration to generate measurable differences; (ii) the *Overall_Performance* metric may not be sensitive to the specific competencies each strategy targets; or (iii) learner-level heterogeneity (motivation, prior knowledge, cultural background) may overwhelm any strategy-specific effect.

6.2 Feature Importance and Explainability: All three ensemble regressors (Gradient Boosting, Random Forest, and Extra Trees) produced R^2 values at or below zero, indicating that the models perform no better and in some cases slightly worse than simply predicting the sample mean for every observation. MAE values of approximately 15 and RMSE values near 17.3 reflect substantial unexplained variability. The binary classification results (ROC-AUC $\approx$ 0.50) confirm this finding from a discriminative perspective: the feature set cannot reliably separate high from low performing learners.

6.3 Feature Importance and SHAP Insights: Despite the overall weak predictive signal, the Gradient Boosting model's feature-importance ranking and SHAP analyses identify a consistent hierarchy of relatively stronger predictors. *Communication_Effectiveness* was the most influential feature within the fitted Gradient Boosting model; however, given the model's near-zero predictive performance, this ranking should be interpreted as model dependent rather than as evidence of a strong causal or predictive relationship with learner performance. It is followed by *Pronunciation_Score, Assignment_Marks, Feedback_Score, and Grammar_Score*. This ordering aligns with theoretical expectations: communicative and linguistic competencies are more proximal predictors of overall English performance than demographic attributes or raw technology-usage counts. However, even the top-ranked feature contributes only a marginal share of explained variance, underscoring the diffuse and multifactorial nature of language acquisition.

6.4 PDP and ICE Analyses: The Partial Dependence and Individual Conditional Expectation plots for *LMS_Login_Count, Discussion_Post_Count*, and *Listening_Score* reveal largely flat average marginal effects, indicating that increases in these variables are not associated with substantial changes in predicted *Overall_Performance* at the population level. The ICE curves, however, display considerable heterogeneity among individual learners, suggesting that the relationship between digital engagement and performance may be moderated by unobserved learner characteristics not captured in the dataset.

6.5 Implications for Intelligent English Learning Systems: The findings highlight a critical gap between the richness of IoT-collected data and its current predictive utility for overall performance. While IoT enabled classrooms generate extensive multimodal records, the present feature set appears insufficient to model the complex, nonlinear, and culturally mediated processes underlying English language acquisition. Future work should consider: (i) incorporating deeper multimodal features (e.g., speech prosody, facial affect, real time interaction dynamics); (ii) employing temporal and sequential modelling to capture learning trajectories rather than single session snapshots; (iii) integrating contextual and cultural moderators; and (iv) exploring alternative target variables (e.g., skill specific gains, communicative competence indices) that may be more amenable to prediction.

7. Conclusion

This study conducted a comprehensive predictive analysis of *Overall_Performance* within an IoT-enabled multicultural English learning dataset comprising 15,000 session records and 54 features. The methodological framework integrated ensemble regression (Gradient Boosting, Random Forest, Extra Trees), binary classification (Random Forest, Logistic Regression), one way ANOVA, and state of the art model interpretation tools (SHAP, PDP/ICE) to examine both predictive accuracy and pedagogical differentiation.

The principal findings are summarised as follows:

1. Teaching-strategy equivalence. One-way ANOVA revealed no statistically significant difference in mean *Overall_Performance* across the six teaching strategies ($F = 0.237$, $p = 0.946$). All group means clustered tightly around 70 with homogeneous standard deviations (~17.3), indicating that no single pedagogical approach demonstrated measurable superiority in this dataset.

2. Negligible regression performance. The three ensemble regressors yielded R^2 values at or below zero (best: Gradient Boosting, $R^2 = -0.0022$), with MAE ≈ 15.0 and RMSE ≈ 17.3. These results indicate that the available predictors explain essentially no variance in *Overall_Performance* beyond the grand mean.

3. Chance-level binary classification. Both Random Forest (ROC-AUC = 0.5047) and Logistic Regression (ROC-AUC = 0.5098) produced discrimination metrics indistinguishable from random guessing for the dichotomised high-performance outcome (≥ 75).

4. Consistent feature hierarchy. Despite weak overall predictivity, feature-importance and SHAP analyses consistently identified *Communication_Effectiveness, Pronunciation_Score, Assignment_Marks, Feedback_Score, and Grammar_Score* as the relatively strongest predictors, confirming the primacy of linguistic and communicative variables over demographic or technology usage features.

5. Flat marginal effects with individual heterogeneity. PDP and ICE plots for selected engagement and linguistic variables showed minimal average marginal effects but substantial between learner variability, suggesting that unmeasured moderators may govern individual learning trajectories.

Finally, these results indicate that while the IoT-enabled dataset captures a rich array of learner characteristics, the current feature set and modelling approaches are insufficient to predict *Overall_Performance* with practical accuracy or to differentiate among teaching strategies. The findings underscore the need for richer multimodal representations, longitudinal modelling frameworks, and alternative outcome metrics to unlock the full potential of AI-driven personalised English education. Nevertheless, the interpretability analyses provide a transparent foundation for future iterative model development and hypothesis generation in technology-enhanced language learning research.

References

[1] Bian, R. (2024). Research on business English translation teaching in the context of economic integration.

[2] Yin, M. (2024). Developing digital tools for cross-cultural adaptability assessment of international students. *In Proceedings of the 2024 International Conference on Big Data Mining and Information Processing* (p. 178–182).

[3] Lin, F. (2022). Internet of things based Korean cross-cultural communication interactive talent training model under curriculum, ideology, and politics. *Scientific Programming, 2022*(1), Article 9122934.

[4] Farida, F., Supardi, S., Abduh, A., Muchtar, J., Rosmaladewi, R., Arham, M. (2024). Technology and hybrid multimedia for language learning and cross-cultural communication in higher education. *ASEAN Journal of Science and Engineering, 4*(2), 331–348.

[5] Long, J., Lin, J. (2022). An empirical study on cultivating college students' cross cultural communicative competence based on the artificial intelligence English teaching mode. *Frontiers in Psychology, 13*, Article 976310.

[6] Richardson, J. W., Khawaja, S. (2025). Meta-synthesis of school leadership competencies to support learner-centered, personalized education. *Frontiers in Education, 10*, Article 1537055. https://doi.org/10.3389/feduc.2025.1537055.

[7] Naseer, F., Khan, M. N., Tahir, M., Addas, A., Aejaz, S. H. (2024). Integrating deep learning techniques for personalized learning pathways in higher education. *Heliyon, 10*(11), Article e32628. https://doi.org/10.1016/j.heliyon.2024.e32628.

[8] Garcia-Ponce, E. E., Tavakoli, P. (2022). Effects of task type and language proficiency on dialogic performance and task engagement. *System, 105*, Article 102734. https://doi.org/10.1016/j.system.2022.102734.

[9] Hashim, H. U. (2024). Globalization, cultural competence, and English language education 5.0: Policy implications and strategies. In *Preconceptions of Policies, Strategies, and Challenges in Education 5.0* (p. 1–16). IGI Global. https://doi.org/10.4018/979-8-3693-3041-8.ch001.

[10] Dong, W., Pan, D., Kim, S. (2024). Exploring the integration of IoT and generative AI in English language education: Smart tools for personalized learning experiences. *Journal of Computer Science, 82*, Article 102397.

[11] El-Mougy, A., Al-Shiab, I., Ibnkahla, M. (2019). Scalable personalized IoT networks. *Proceedings of the IEEE, 107*, 695–710.

[12] Kamruzzaman, M. M., Alanazi, S., Alruwaili, M., Alshammari, N., Elaiwat, S., Abu-Zanona, M., Innab, N., Mohammad Elzaghmouri, B., & Ahmed Alanazi, B. (2023). AI- and IoT-assisted sustainable education systems during pandemics, such as COVID-19, for smart cities. *Sustainability, 15*, Article 8354.

[13] Sharef, N. M., Azmi Murad, M. A., Mansor, E. I., Nasharuddin, N. A., Omar, M. K., Samian, N., Arshad, N. I., Ismail, W., Shahbodin, F. (2020, October 20–21). Learning-analytics based intelligent simulator for personalised learning. *In Proceedings of the 2020 International Conference on Advancement in Data Science, E-learning and Information Systems (ICADEIS)* (pp. 1–6). Lombok, Indonesia.

[14] Truman, B.,Truman, J. M. (2017). Advancing personal learning using the Internet of Things: Creating bonds for societal inclusivity. In *Integrating an Awareness of Selfhood and Society into Virtual Learning* (pp. 258–274). IGI Global.

[15] Huiying, X., Qiang, M. (2021). College English cross cultural teaching based on cloud computing MOOC platform and artificial intelligence. *Journal of Intelligent & Fuzzy Systems, 40*(4), 7335–7345.

[16] Zhou, L., Sun, X., Mu, G., Wu, J., Zhou, J., Wu, Q., et al. (2021). A tool to facilitate the cross-cultural design process using deep learning. *IEEE Transactions on Human-Machine Systems, 52*(3), 445–457.

[17] Alkato, M. A., Kalenen, N. (2025). Dynamic pattern analysis for enhanced predictive intelligence in smart environments using transformer learning models. *PatternIQ Mining, 2*(1), 85–96. https://doi.org/10.70023/sahd/250208.

[18] Dai, H., Zhao, T. (2022). Intelligent analysis strategy of pragmatic failure in cross cultural communication based on convolution neural network. *Mobile Information Systems, 2022*(1), Article 7803959.

[19] Jiang, P., Hou, X. (2022). Development and application of MOOC system for English intercultural communication courses using neural network. *Scientific Programming, 2022*(1), Article 9882113.

[20] Huiying, X., Qiang, M. (2021). College English cross-cultural teaching based on cloud computing MOOC platform and artificial intelligence. *Journal of Intelligent & Fuzzy Systems, 40*(4), 7335–7345.

[21] Zubair, A., Al Rashed, F. (2024). Deep learning algorithms for multimodal interaction using speech and motion data in virtual reality systems. *PatternIQ Mining, 1*(4), 52–64. https://www.doi.org/10.70023/sahd/241105.

[22] Han, Q. (2024). Research on the accuracy of machine translation in cross-cultural communication based on embedded neural networks. *International Journal of High Speed Electronics and Systems*, Article 2540125.

[23] Ma, D. (2026). Optimization of English cross-cultural communication ability improvement method based on deep learning and Internet of Things. *Discover Artificial Intelligence, 6*, Article 317. https://doi.org/10.1007/s44163-026-01007-7.

[24] Zhu, X. (2026). Integrating English learning with virtual reality and the Internet of Things to enhance the educational experience. *International Journal of High Speed Electronics and Systems, 35*(2), Article 2550003.

[25] Zheng, H., Zhong, D. (2026). Optimization of methods for improving English cross-cultural communication skills based on deep learning and the internet of things. *Discover Internet of Things*.

[26] Tian, J., Liu, Z. (2026). Design of a personalized English learning recommendation system based on AI and IoT. *Discover Internet of Things, 6*, Article 85. https://doi.org/10.1007/s43926-026-00355-1.

[27] Spaho, E., Çiço, B., Shabani, I. (2025). IoT integration approaches into personalised online learning: Systematic review. *Computers, 14*(2), Article 63.

[28] Liu, Z. (2026). Design and effect evaluation of college English software teaching platform based on IoT integration. *International Journal of High Speed Electronics and Systems, 35*(2), Article 2540569.

[29] Kaisar, S., Chowdhury, A. (2022). Integrating oversampling and ensemble-based machine learning techniques for an imbalanced dataset in dyslexia screening tests. *ICT Express, 8*(4), 563–568.

[30] Lv, H., Yan, K., Guo, Y., Zou, Q., Hesham, A. E., Liu, B. (2022). AMPpred-EL: An effective antimicrobial peptide prediction model based on ensemble learning. *Computers in Biology and Medicine, 146*, Article 105577. PMID: 35576825.

[31] Sun, J., Wu, S., Zhang, H., Zhang, X., Wang, T. (2022). Based on multi algorithm hybrid method to predict the slope safety factor – Stacking ensemble learning with Bayesian optimization. *Journal of Computational Science, 59*, Article 101587.

[32] Tong, T., Li, Z. (2025). Predicting learning achievement using ensemble learning with result explanation. *PLoS ONE, 20*(1), Article e0312124. https://doi.org/10.1371/journal.pone.0312124.

[33] Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of Statistics, 24*, 303–329.

[34] Park, J., Lee, W. H., Kim, K. T., Park, C. Y., Lee, S., Heo, T. Y. (2022). Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. *Science of the Total Environment, 832*, Article 155070.

[35] Yang, Z., Sudjianto, A., Li, X., Zhang, A. (2024). Inherently interpretable tree ensemble learning. *arXiv preprint arXiv:2410.19098.*

[36] Wang, W., Zhao, Y., Li, Y. (2024). Ensemble machine learning for predicting the homogenized elastic properties of unidirectional composites: A SHAP-based interpretability analysis. *Acta Mechanica Sinica, 40*, Article 423-301. https://doi.org/10.1007/s10409-023-23301.

[37] Kaggle. (n.d.). *English learning interaction dataset* [Data set]. https://www.kaggle.com/datasets/colabsss/english-learning-interaction-dataset.