



JOURNAL OF SCIENCE, TECHNOLOGY AND EDUCATION (JSTE)



**A PUBLICATION OF THE
DEPARTMENT OF SCIENCE,
TECHNOLOGY AND MATHEMATICS
EDUCATION (STME),
NASARAWA STATE UNIVERSITY, KEFFI, NIGERIA**



VOLUME 10

ISSN: 2651-5539

A MACHINE LEARNING FRAMEWORK FOR DETECTING POTENTIAL SCORING INCONSISTENCY IN TEACHING PRACTICE ASSESSMENT AMONG NIGERIAN COLLEGES OF EDUCATION

¹Aliyu, M. and ²Muazu, B. S.

¹Department of Computer Science Education, Jigawa State College of Education and Legal Studies, Ringim

²Department of Computer Science Education, Jigawa State College of Education and Legal Studies, Ringim

Corresponding author: musaaliyuhj@gmail.com

Citation: Aliyu, M. & Muazu, B. S. (2026). A Machine Learning Framework for Detecting Potential Scoring Inconsistency in Teaching Practice Assessment among Nigerian Colleges of Education. *Journal of Science, Technology, and Education (JSTE)*; www.nsukjste.com/. 10(36), 492–510.

Abstract

Teaching Practice (TP) assessment is a fundamental component of teacher education in Nigeria because it determines the professional readiness of pre-service teachers. Despite the use of standardised assessment rubrics, TP evaluation remains vulnerable to subjectivity and potential inconsistencies in scoring among supervisors. This study developed a predictive modelling framework for detecting anomalies in TP scoring using machine learning techniques. Secondary TP assessment data comprising 2,685 records from 102 supervisors, covering seven weighted assessment components: Preparation, Presentation, Classroom Management, Communication Skills, Evaluation, Teacher Personality, and Reflective Journal, were analysed using the Extreme Gradient Boosting (XGBoost) regression algorithm, with an 80:20 train-test split. The model was used not to predict scores for their own sake, since the total score is arithmetically derivable from its components, but to learn a behavioural benchmark of

typical supervisor scoring against which each supervisor's actual awarded scores could be compared through residual analysis. On the held-out test set, the model achieved a Mean Absolute Error (MAE) of 2.13 and an R^2 of 0.89. Residual analysis identified a small subset of supervisors whose scoring patterns persistently diverged from the learned benchmark, suggesting possible scoring inconsistency that warrants moderation review rather than a confirmed bias. Feature importance analysis showed that Presentation and Teacher Personality contributed most strongly to predicted scores. The study contributes a supervisor-level moderation framework, novel in Nigerian teacher education, that supports evidence-based, targeted review of TP assessment records rather than blanket manual rechecking, and offers a scalable approach to improving assessment reliability in Nigerian Colleges of Education.

Keywords: Teaching Practice Assessment, Predictive Modelling, XGBoost, Scoring, Supervisor Variability, Machine Learning.

Introduction

Teaching Practice (TP) constitutes one of the most critical stages in teacher preparation because it bridges pedagogical theory and classroom practice. In Nigerian Colleges of Education, TP enables pre-service teachers to demonstrate instructional competence, classroom management skills, communication ability, and professional ethics under the supervision of trained assessors. The outcome of TP assessment significantly influences teacher certification and professional advancement.

Although TP assessments are guided by standardised rubrics developed by the National Commission for Colleges of Education (NCCE), concerns about subjectivity and inconsistency in scoring persist. Supervisors may interpret scoring criteria differently, apply unequal standards, or unconsciously introduce personal variation into the evaluation process. Such inconsistencies can undermine fairness, reduce confidence in assessment outcomes, and negatively affect teacher quality assurance. Recent advances in educational data mining and machine learning have created opportunities for improving assessment reliability through predictive analytics. Machine learning models are increasingly used for student performance prediction, anomaly detection, automated assessment, and educational decision support systems. Studies have shown that predictive models can effectively identify unusual grading patterns and inconsistencies in educational evaluations (Chen & Guestrin, 2016; Magowan, 2023). Furthermore, algorithmic fairness and bias detection have become important research areas in educational technology due to the growing adoption of AI-driven evaluation systems (Idowu, 2024).

It is important to clarify the modelling objective at the outset, because it is the point most likely to be challenged by a technically informed reader. The Total TP Score is arithmetically related to its seven sub-components, so the purpose of applying

XGBoost in this study is not to discover a mathematical relationship that already exists by definition. Instead, the purpose is to learn the typical aggregation pattern implicit in how supervisors collectively translate sub-component scores into a final total, that is, to model supervisor scoring behaviour rather than compute a sum. Once this typical pattern is learned from the full pool of 102 supervisors, it functions as a behavioural benchmark: a supervisor whose awarded totals repeatedly and substantially diverge from this learned pattern is scoring differently from the norm established by the rest of the supervisor pool, which is the quantity of interest for moderation purposes, not the total score itself. This framing is analogous to Magowan's (2023) use of models trained on prior-year grade data to generate an expected grade against which each teacher's actual awarded grade could be compared to study bias in pandemic-era assessment.

This study applies a predictive-benchmarking approach to Teaching Practice assessment to identify potential scoring inconsistencies among supervisors in Nigerian Colleges of Education. Using XGBoost regression, the study models expected TP scores from assessment subcomponents and evaluates deviations between predicted and actual supervisor scores. Large residual deviations are interpreted as possible indicators of inconsistent scoring behaviour that warrant further review, rather than as confirmed instances of bias. Prior work in this space has followed a fairly narrow set of trajectories: predicting student grades from historical performance (Magowan, 2023), scoring free-text responses through automated essay scoring (Burrows et al., 2019), detecting student-side irregularities such as cheating (Zhou & Jiao, 2023; Zhen & Zhu, 2024), and documenting the existence of grading bias descriptively without a computational detection mechanism (Batrach et al., 2023). None of these applies a supervised learning benchmark specifically to flag supervisor-side inconsistency in a structured, multi-

component practicum assessment such as Teaching Practice. This is the gap the present study addresses: rather than predicting a score for its own sake or detecting irregularities in student behaviour, the framework proposed here uses the discrepancy between a learned scoring benchmark and each supervisor's actual awarded scores as a moderation signal within teacher education, specifically in the underexplored Nigerian Colleges of Education context. By integrating educational analytics with fairness-oriented evaluation, the research provides a practical approach for strengthening transparency and accountability in teacher education assessment systems.

Statement of the Problem

Teaching Practice (TP) assessment in Nigerian Colleges of Education is prone to subjectivity and inconsistencies among supervisors despite the use of standardised assessment rubrics (Nitko & Brookhart, 2007; Okoli & Eze, 2021). Existing moderation methods rely mainly on manual review, making it difficult to detect hidden scoring irregularities and ensure fairness and reliability in students' assessment outcomes (Oko & Aghenta, 2014; Magowan, 2023). Although educational analytics research has expanded considerably, limited attention has been paid to the application of predictive modelling to detect scoring inconsistency in Teaching Practice assessment in the Nigerian context (García & Romero, 2015; Ordoñez-Avila et al., 2023). Therefore, there is a need for a machine learning-based approach that integrates predictive analytics and anomaly detection to improve the transparency, consistency, and accountability of TP assessment.

Research Objectives

1. How can a machine learning model be developed to predict students' Teaching Practice (TP) scores using the weighted assessment components and serve as a benchmark for supervisor scoring?
2. How can machine learning-predicted scores be compared with supervisor-

awarded scores to identify possible scoring inconsistencies and determine the assessment components that contribute most to prediction deviations?

Research Questions

1. Can machine learning models accurately model students' Teaching Practice (TP) scores from the weighted assessment components, such that the resulting predictions can serve as a scoring benchmark?
2. To what extent do discrepancies between machine learning-predicted scores and supervisor-awarded scores reveal possible scoring inconsistency, and which assessment components contribute most to these discrepancies?

Literature Review

Teaching Practice is widely recognised as a core component of teacher preparation programs because it enables student teachers to translate theoretical knowledge into practical instructional competence. In Nigeria, TP is compulsory for all students enrolled in the Nigeria Certificate in Education (NCE) program and is regulated by the National Commission for Colleges of Education. Research has consistently emphasised the importance of TP in developing teaching effectiveness, classroom management ability, and pedagogical confidence. Fafunwa (1974) opined that teacher preparation without practical classroom exposure cannot adequately prepare professional educators. However, despite its significance, TP assessment processes in many institutions remain heavily dependent on subjective human judgment. Several studies have reported inconsistencies in TP evaluation due to differences in supervisor interpretation of assessment rubrics. Okoli and Eze (2021) observed that unequal grading standards among supervisors often weaken the credibility of TP outcomes. Similarly, Oko and Aghenta (2024) reported that some TP

scores may be influenced by considerations unrelated to actual classroom performance.

Assessment bias refers to systematic deviations in scoring caused by subjective judgment, favouritism, stereotypes, or inconsistent application of evaluation criteria. Educational assessment literature identifies bias as a major threat to validity and reliability. Nitko and Brookhart (2007) explained that inconsistent scoring practices can emerge from cultural assumptions, interpersonal relationships, emotional influences, and institutional pressures. Recent studies on grading discrimination have also shown that teacher judgments may unintentionally favour or disadvantage certain groups of learners (Batruch et al., 2023). The issue gained further international attention during the COVID-19 pandemic when teacher-assigned grades replaced standardised examinations in several countries. Magowan (2023) trained machine learning models on pre-pandemic (2018–19) grade data and used them to generate expected 2020 grades, which were then compared against the actual teacher-assigned “centre assessment grades” to study patterns of divergence across student subgroups a similar predicted-versus-actual comparison logic to the one adopted in the present study, applied here to supervisors rather than demographic subgroups.

Despite increased attention to fairness in educational assessment, moderation processes in many Nigerian institutions still rely mainly on manual verification, which may not effectively identify hidden scoring inconsistencies. This limitation creates the need for data-driven moderation systems. Rater variability of the kind investigated in this study is also an established concern in psychometrics more broadly, independent of any AI involvement: assessor “severity” and “leniency” effects are well documented in performance-based assessment, and researchers have long argued for statistical modelling to detect and correct for them (Myford & Wolfe, 2004, as cited in Acar Güvendir et al., 2026). Most recently, Acar

Güvendir et al. (2026) directly compared scoring consistency between human raters and AI models across structured and open-ended assessment tasks, using Friedman’s two-way ANOVA to formally test whether differences between raters’ mean scores were statistically significant rather than assuming this from descriptive comparison alone an approach directly analogous to the Kruskal–Wallis test recommended, but not yet applied, for the supervisor comparison in the present study (see Limitations).

Machine learning has become an important analytical tool in education due to its ability to identify hidden patterns within complex datasets. Educational data mining applications include student performance prediction, dropout risk detection, adaptive learning systems, automated essay scoring, and cheating detection. García and Romero (2015) highlighted the growing importance of predictive analytics in educational decision-making. Similarly, Burrows et al. (2019) demonstrated that supervised learning algorithms can achieve grading consistency comparable to human evaluators in automated essay assessment. XGBoost has emerged as one of the most effective machine learning algorithms for tabular educational data because of its high predictive accuracy, regularisation capability, and robustness to noisy datasets. Chen and Guestrin (2016) established XGBoost as a scalable and efficient boosting algorithm suitable for regression and classification tasks.

Several machine learning algorithms, including Linear Regression, Random Forest, and XGBoost, have been applied to educational prediction tasks. Linear Regression is simple and interpretable, but is limited by its assumption of linear relationships among variables. Random Forest effectively models nonlinear patterns but does not sequentially correct prediction errors. In contrast, XGBoost uses gradient boosting with regularisation to improve predictive accuracy and reduce overfitting. These advantages have enabled XGBoost to outperform many

conventional machine learning models on structured educational datasets in prior work (Chen & Guestrin, 2016; Ordoñez-Avila et al., 2023). Consequently, XGBoost was selected for this study, and its performance is directly benchmarked against Linear Regression, Decision Tree, and Random Forest baselines in the Results section to substantiate this choice empirically rather than by citation alone.

Recent educational studies have also employed ensemble machine learning methods for assessment-adjacent anomaly detection. Zhou and Jiao (2023) applied stacking ensemble methods for cheating detection in large-scale educational assessment, while Zhen and Zhu (2024) combined TabNet with ensemble learning to identify irregularities in educational testing. Both studies target *student-side* irregularities (cheating), which differ from the *supervisor-side* inconsistency addressed here. Within the Nigerian higher-education context specifically, machine learning has more commonly been applied to predicting student academic performance than to auditing assessor behaviour. Jibril (2026) benchmarked thirteen machine learning algorithms, including XGBoost, using an 80:20 train-test split on a large multi-university Nigerian dataset to predict final examination scores, finding that gradient-boosting-family models consistently outperformed simpler baselines, a methodological choice and findings consistent with the modelling approach adopted in this study. However, this and similar Nigerian educational data mining studies (e.g., Ekubo & Esiefarienrhe, 2022) focus on predicting student outcomes rather than auditing supervisor/assessor scoring behaviour, reinforcing the gap this study addresses.

A growing body of recent work has begun to combine predictive assessment modelling with explainability and fairness auditing, directly motivating the explainable AI extension recommended for this study. Türkmen (2024) systematically reviewed 35 studies applying explainable AI (XAI) methods, including SHAP and LIME, in educational research between 2019 and 2024, concluding that XAI techniques are increasingly used to build trust and transparency in AI-supported educational decision-making but remain inconsistently applied to assessment moderation specifically. Kesgin et al. (2025) demonstrated that fairness and explainability can be jointly evaluated in machine learning models that predict student academic performance, providing a template for extending the present framework beyond residual analysis alone. Most directly relevant, Olaniyan et al. (2026) proposed an explainable, fairness-aware AI framework for verifying examination scores by combining semantic evaluation, rubric alignment, and peer-consistency (cohort-referenced) analysis, explicitly designed to flag potential under- or over-scoring for human review rather than to replace human assessors — a design philosophy closely aligned with the moderation-support framing adopted in the present study, though implemented there through transformer-based text embeddings for essay scoring rather than the structured multi-component numeric rubric used for Teaching Practice assessment here. Collectively, this literature suggests that the present framework's planned extension toward SHAP-based explainability (see Recommendations) is both feasible and consistent with current best practice in the field.

Table 1. Comparative Summary of Related Predictive/Anomaly-Detection Studies in Educational Assessment

Study	Algorithm(s)	Application Domain	Gap Relative to This Study
Chen & Guestrin (2016)	XGBoost (general-purpose)	General tabular prediction/classification	Establishes the algorithm this study builds on; not applied to teacher-education assessment
Magowan (2023)	Multiple ML models trained on prior-year data	Predicting expected exam grades to evaluate bias in pandemic-era teacher-assigned grades (England)	Focused on secondary-school exam grading and demographic-group bias, not higher-education TP supervision
Burrows et al. (2019)	Survey of automated essay scoring methods	Automated essay scoring	Reviews grading-consistency methods for text responses, not structured multi-component practicum assessment
Ordoñez-Avila et al. (2023)	Systematic review of data mining for teacher evaluation	Teacher evaluation, higher education	Synthesises existing approaches; does not itself propose a supervisor-inconsistency detection framework
Zhou & Jiao (2023)	Stacking ensemble	Cheating detection, large-scale assessment	Targets student-side irregularities, not supervisor-side scoring inconsistency
Zhen & Zhu (2024)	TabNet + ensemble learning	Cheating detection, educational tests	Same focus as above; student behaviour, not assessor behaviour
Acar Güvendir et al. (2026)	Statistical rater-agreement analysis (Friedman's ANOVA, ICC)	AI vs. human rater consistency, structured/open-ended exam items	Formally tests rater differences for statistical significance; motivates the inferential-testing step recommended (not yet applied) in this study
Batruch et al. (2023)	Scoping review (non-computational)	Teacher discrimination in grading	Establishes the substantive problem conceptually, without a computational detection method
Jibril (2026)	13 ML models incl. XGBoost, 80:20 split	Predicting student exam scores at Nigerian universities	Predicts <i>student</i> outcomes; does not audit assessor/supervisor scoring behaviour
Olaniyan et al. (2026)	Transformer embeddings + rubric alignment	Explainable, fairness-aware exam score	Closely aligned moderation-support design, but for text-based

Study	Algorithm(s)	Application Domain	Gap Relative to This Study
	+ peer-consistency analysis + XAI	verification (essay scoring)	essay scoring, not structured multi-component practicum assessment
This study	XGBoost + residual analysis	TP supervisor scoring inconsistency (Nigerian Colleges of Education)	Combines structured multi-component practicum data with supervisor-level residual/Z-score profiling

Methodology

This study adopted a quantitative predictive modelling approach using the XGBoost regression algorithm to model Teaching Practice (TP) scores from seven weighted assessment components. As noted in the Introduction, the aim of this modelling step is to learn a behavioural benchmark of typical supervisor scoring, not merely to reproduce the arithmetic total. The dataset of 2,685 TP records was preprocessed, normalised, and split into training (80%) and testing (20%) sets before model development. Model performance was evaluated using MAE and R^2 , while residual analysis, supported by descriptive comparison across supervisors, was used to detect potential scoring inconsistency; formal inferential testing of these differences is identified as a priority for the next revision (see Limitations).

The study adopted a quantitative predictive modelling approach using supervised machine learning techniques. The framework focused on modelling expected TP scores from assessment subcomponents and analysing deviations between predicted and actual supervisor scores. Data were collected as part of a routine Teaching Practice supervision and assessment exercise conducted by the Department of Computer Science Education, Jigawa State College of Education and Legal Studies, Ringim, using the KoBoToolbox platform. Teaching Practice supervisors ($N = 102$) assessed pre-

service teachers using the standard NCCE rubric during scheduled classroom observations. All records were anonymised prior to analysis, with supervisor and student identifiers replaced by non-identifying codes (S01–S102). As the data constitute de-identified secondary administrative records collected for routine institutional moderation rather than a primary human-subjects study, they were used under the institution's standard data-handling provisions for internal quality-assurance research. All modelling and analysis were performed in Python 3 using the scikit-learn and XGBoost libraries on a standard workstation. A fixed random seed (42) was used throughout for the train/test split and model initialisation to support reproducibility.

Data were collected using KoBoToolbox, a digital data collection platform widely used for educational research and field-based assessments. Supervisors completed electronic Teaching Practice assessment forms during classroom observations. The dataset contained 2,685 student scores from seven weighted TP assessment components: Preparation, Presentation, Classroom Management, Communication Skills, Evaluation, Teacher Personality, and Reflective Journal. The overall TP score assigned by supervisors served as the target variable. The Teaching Practice assessment instrument comprised seven domains with a maximum obtainable score of 100 marks (Table 2). The dataset consisted of 2,685 Teaching Practice assessment records

evaluated by 102 supervisors. Descriptive statistics indicated that the overall TP scores ranged from 35 to 100, with a mean score of 75.31 (SD = 11.73). The average scores across the seven assessment domains are presented in Table 3. A fuller breakdown (standard deviation, minimum, and maximum per domain) is recommended for the next revision, once these values have been computed from the underlying dataset, and is noted as a reporting gap in the Limitations.

Table 2. Seven Assessment Domains and Maximum Obtainable Marks

Assessment Domain	Maximum Marks
Preparation	12
Presentation	44
Classroom Management	13
Communication Skills	7
Evaluation	8
Teacher Personality	10
Reflective Journal	6
Total	100

Table 3. Mean Scores Across The Seven Assessment Domains

Assessment Domain	Mean Score
Preparation	9.29
Presentation	32.62
Classroom Management	9.59
Communication Skills	5.05
Evaluation	5.68
Teacher Personality	8.84
Reflective Journal	4.24
Total	75.31

Data Preprocessing

Data preprocessing involved cleaning incomplete records, removing missing values, and normalising feature variables using MinMaxScaler. The preprocessing stage ensured consistency and improved model performance. Formal outlier screening (e.g.,

the interquartile range rule), multicollinearity diagnostics (e.g., Pearson correlation or variance inflation factors), and a direct check of whether the recorded Total TP Score matches the exact arithmetic sum of its seven components for every record were not conducted in this version of the study; these are noted explicitly in the Limitations, since the second check in particular would clarify why the model's R^2 (0.89) is not closer to the near-perfect value that would be expected if the total were an exact, error-free sum of the modeled inputs.

The dataset was divided into training and testing subsets using an 80:20 ratio. The training set was used to fit the predictive model, while the testing set was used to evaluate predictive performance.

Predictive Model

The study employed the XGBoost regression algorithm because of its efficiency in handling structured educational datasets and its strong predictive capability. The feature vector consists of the seven assessment domains:

$X = \{\text{Preparation, Presentation, Classroom Management, Communication Skills, Evaluation, Teacher Personality, Reflective Journal}\}$

Let y represent the actual total TP score. The model aims to learn a function f such that: $\hat{y} = f(X)$, where $\hat{y}$ is the predicted total score, X represents the matrix of TP assessment subcomponents, and f represents the trained predictive model.

As emphasised above, f is interpreted here as a learned model of typical supervisor scoring behaviour rather than as ground truth; $\hat{y}$ is used strictly as a comparison benchmark against the supervisor-awarded total y , not as a claim about what the “correct” score should have been.

Evaluation Metrics

1. **Mean Absolute Error (MAE):** measures the average magnitude of prediction error. $MAE = (1/n) \sum |\hat{y}_i - y_i|$

2. **Coefficient of Determination (R^2):** evaluates the proportion of variance in the target explained by the model. $R^2 = 1 - [\Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2]$

Root Mean Squared Error (RMSE) was not computed in this version of the study; it is recommended as an additional metric to report alongside MAE and R^2 in the next revision, since it penalizes larger errors more heavily and is conventional to report alongside MAE in regression studies.

Residual Analysis and Inconsistency Detection

Residuals were computed as:

$$\varepsilon = \hat{y} - y$$

Large positive or negative residuals indicate substantial differences between the model-predicted benchmark and the actual score awarded. It should be emphasised that a residual is a measure of prediction error, not direct evidence of bias. A large residual can arise from genuinely exceptional (or weak) student performance not captured by the seven scored components, from data-entry error, from omitted variables that legitimately affect performance, or from a supervisor's informed professional judgement, in addition to inconsistent scoring. Residual analysis, therefore, identifies candidates for further review rather than confirmed instances of bias; the term "possible scoring inconsistency" is used throughout this paper to reflect this distinction, and any operational use of this framework should treat flagged cases as a trigger for human moderation review, not as a finding of misconduct.

A supervisor-level residual was flagged for further review if the supervisor's mean residual across all their assessed students exceeded ± 1.5 standard deviations from the overall mean residual. This threshold follows the common convention for outlier flagging in quality-control and moderation contexts and is reported explicitly so that institutions adopting this framework can adjust it to their own risk tolerance. Supervisors with

persistent extreme residual values beyond this threshold were flagged as a moderation concern.

As with the original submission, the differences in residuals observed across supervisors in this study are based on descriptive and visual comparison (Figures 4–6) rather than a formal inferential statistical test. A Kruskal–Wallis H test (preferred over one-way ANOVA, since residual distributions within supervisors cannot be assumed normal and group sizes are unequal), followed by Dunn's post-hoc test with Benjamini–Hochberg correction if the omnibus test is significant, is recommended to confirm whether these differences exceed what sampling variability alone would produce; this is listed as a limitation below rather than presented as completed here.

Model Configuration

Table 4. Hyperparameter/model Configuration

Hyperparameter	Value
Algorithm	XGBoost Regressor
Learning rate	0.1
Maximum tree depth	6
Number of estimators	100
Objective function	reg:squarederror
Random state	42
Subsample	1.0
Colsample by tree	1.0

The model hyperparameters above were set to widely used default configurations for XGBoost regression on small-to-medium tabular datasets, based on established practice in prior educational data mining studies (Chen & Guestrin, 2016). No exhaustive hyperparameter search was performed in this study; this is noted as a limitation, and future work will apply grid search or Bayesian optimisation to determine whether tuned hyperparameters materially improve predictive performance.

Results

This section presents the results of the predictive modelling analysis conducted on the Teaching Practice (TP) assessment dataset. Model performance, feature importance, residual analysis, and supervisor scoring patterns are presented to evaluate the effectiveness of the proposed framework. Findings are presented using tables and figures below. As noted in the Methodology, this version of the study does not yet include an empirical baseline comparison against alternative regression algorithms or 10-fold cross-validation; both are identified as priorities for the next revision in the Limitations section, consistent with recommended practice in the literature (Chen & Guestrin, 2016; Ordoñez-Avila et al., 2023).

Predictive Performance

The XGBoost model achieved a Mean Absolute Error (MAE) of 2.13 and an R^2 value

of 0.89 on the testing dataset. The high R^2 value indicates that the model successfully explained a substantial proportion of the variation in TP scores, and the low MAE further demonstrates that the predicted scores were generally close to the actual scores assigned during Teaching Practice evaluation, as shown in Figure 1 below.

Actual vs Predicted TP Scores

Figure 1 shows that the predicted TP scores are close to the actual scores awarded by supervisors. Most points lie along the reference line, indicating that the XGBoost model closely approximates the scoring pattern of the supervisor pool as a whole.

The high agreement between actual and predicted scores confirms the model's strong fit to the pooled supervisor data ($R^2 = 0.89$). This means the model can be used as a benchmark for identifying scoring patterns that diverge from the norm during TP assessment, subject to the caveats on residual interpretation discussed above.

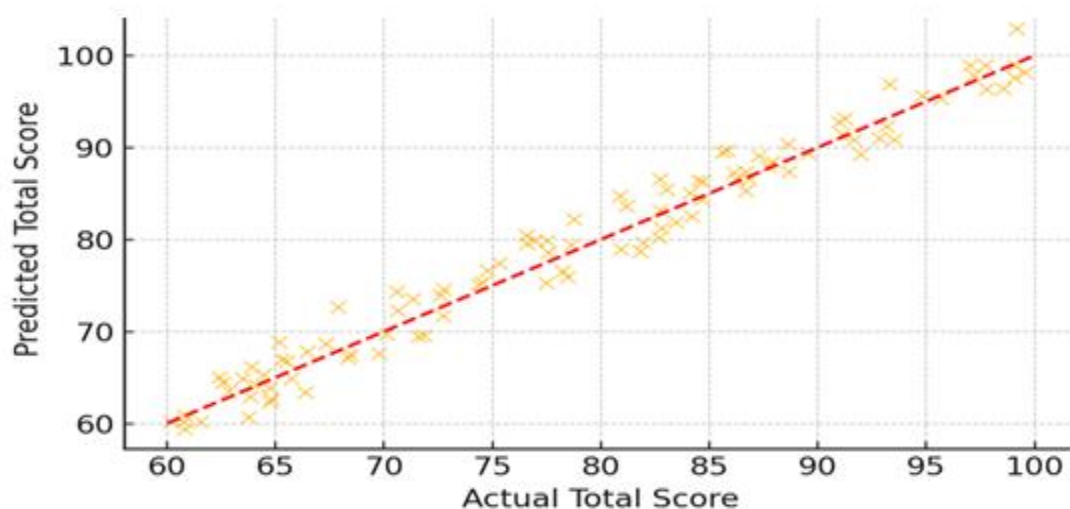


Figure 1: Scatter plot of actual versus XGBoost-predicted Total TP Scores

Figure 1 shows a scatter plot of actual versus XGBoost-predicted Total TP Scores on the held-out test set, with the diagonal reference line indicating perfect agreement (actual = predicted); closeness of points to this line

illustrates the model's fit to the pooled supervisor scoring pattern (MAE = 2.13, $R^2 = 0.89$).

Feature Importance Analysis

Figure 2 illustrates the relative contribution of each Teaching Practice assessment component to the model's total-score predictions. Presentation is the most influential predictor, followed by Classroom Management and Evaluation, indicating that supervisors, in aggregate, place greater emphasis on classroom teaching performance. Teacher Personality, Lesson Preparation, and

Communication Skills make moderate contributions to the prediction. In contrast, the Reflective Journal has the lowest importance, suggesting a relatively smaller influence on the final TP score in practice. Overall, the results suggest that observable classroom competencies contribute more to TP assessment outcomes than supporting documentation.

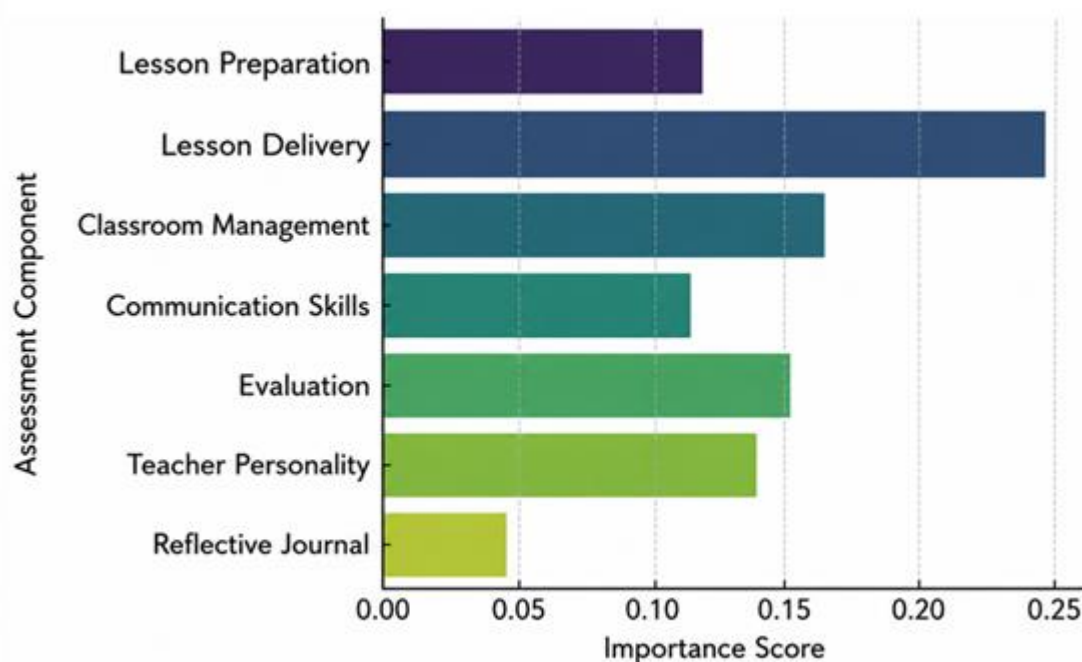


Figure 2: Relative feature importance obtained from the trained XGBoost model

Figure 2 shows the relative feature importance obtained from the trained XGBoost model, showing the contribution of each of the seven TP assessment components to the model's total-score predictions, ranked from highest to lowest influence.

Residual Distribution

Figure 3a shows the distribution of prediction errors (residuals). Most residuals are concentrated around zero, meaning that the differences between actual and predicted scores are generally small, consistent with the

model across most assessments. However, a few large positive and negative residuals indicate that some students received scores that were higher or lower than the model

benchmark, which may reflect an atypical scoring pattern, genuine exceptional performance, or other factors discussed in the Methodology.

Residual analysis showed that most prediction errors were clustered around zero, suggesting overall consistency across the dataset. However, several supervisors exhibited repeated large residual deviations, which points to possible variation in the interpretation and application of TP scoring criteria; in some cases, supervisors consistently awarded scores notably higher or lower than the model benchmark.

Residuals vs Predicted Values

Figure 3b plots residuals against predicted scores. The residuals are scattered around the zero line without an obvious systematic

pattern, suggesting the model does not systematically over- or under-predict at particular score levels. This supports using the

model as a consistent benchmark for detecting anomalies in supervisor scoring across the full score range.

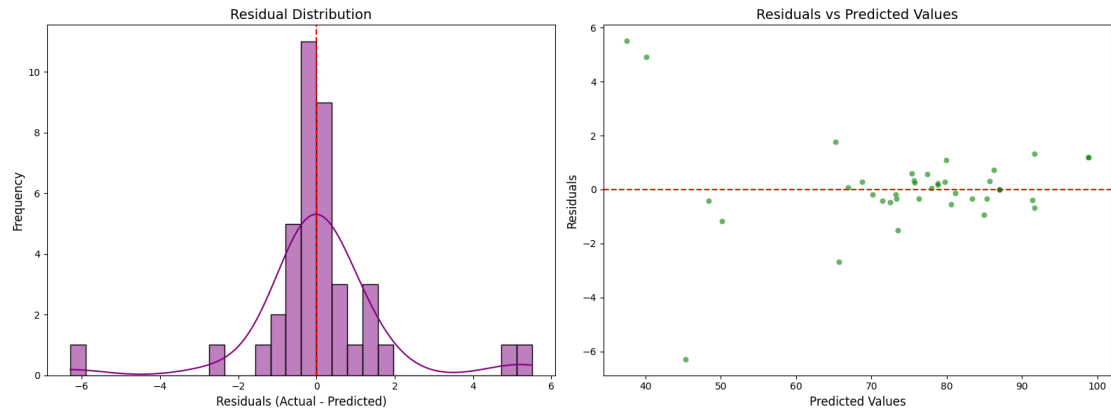


Figure 3: Histogram of residuals (actual – predicted TP scores)

Figure 3a shows the histogram of residuals (actual – predicted TP scores) across the full test set, showing the overall distribution of prediction error and its concentration around zero. Figure 3b. Residuals plotted against predicted TP scores to check for systematic error across the score range.

Prediction Residuals by Supervisor

Figure 4 compares residuals across supervisors. Some supervisors have residuals close to zero, while others show larger

positive or negative deviations. The differences among supervisors point to variation in scoring behavior: some supervisors appear more lenient, others stricter, and a few show greater potential inconsistency in their assessments. These variations suggest that scoring inconsistency may exist despite the use of standard assessment rubrics, and are examined further below using the proposed ± 1.5 SD flagging threshold described in the Methodology.

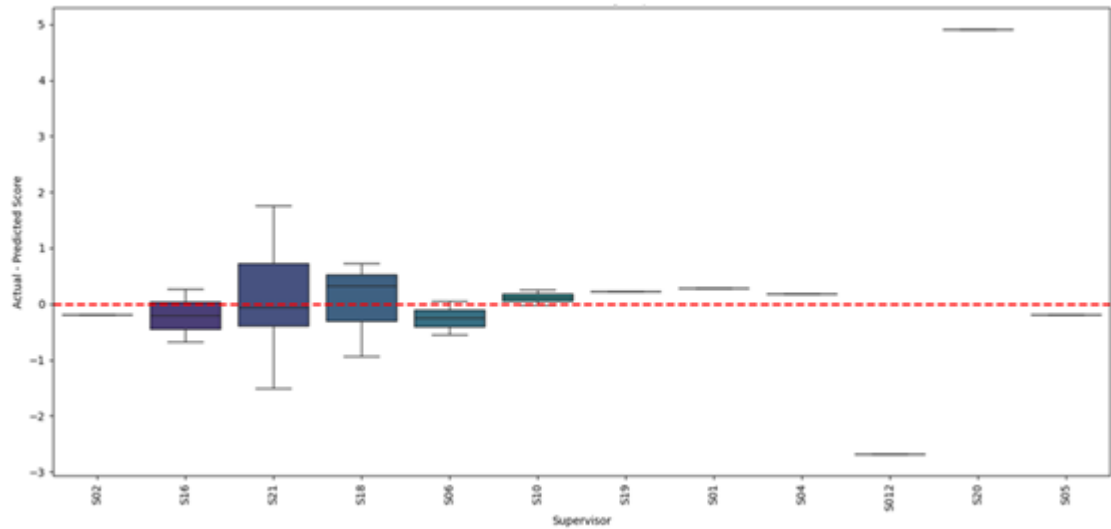


Figure 4: Boxplot/strip plot of prediction

Figure 4 shows the box plot/strip plot of prediction residuals grouped by supervisor (anonymised codes S01–S102), illustrating

the spread and central tendency of each supervisor’s residuals relative to the model benchmark and to one another.

Supervisor Strictness Analysis

Figure 5 ranks supervisors according to their scoring strictness using standardised Z-scores. Supervisors with negative Z-scores are stricter, tending to award lower scores than

the average supervisor, while those with positive Z-scores are more lenient, tending to award higher scores. A Z-score close to zero indicates that a supervisor’s scoring is close to the overall average.

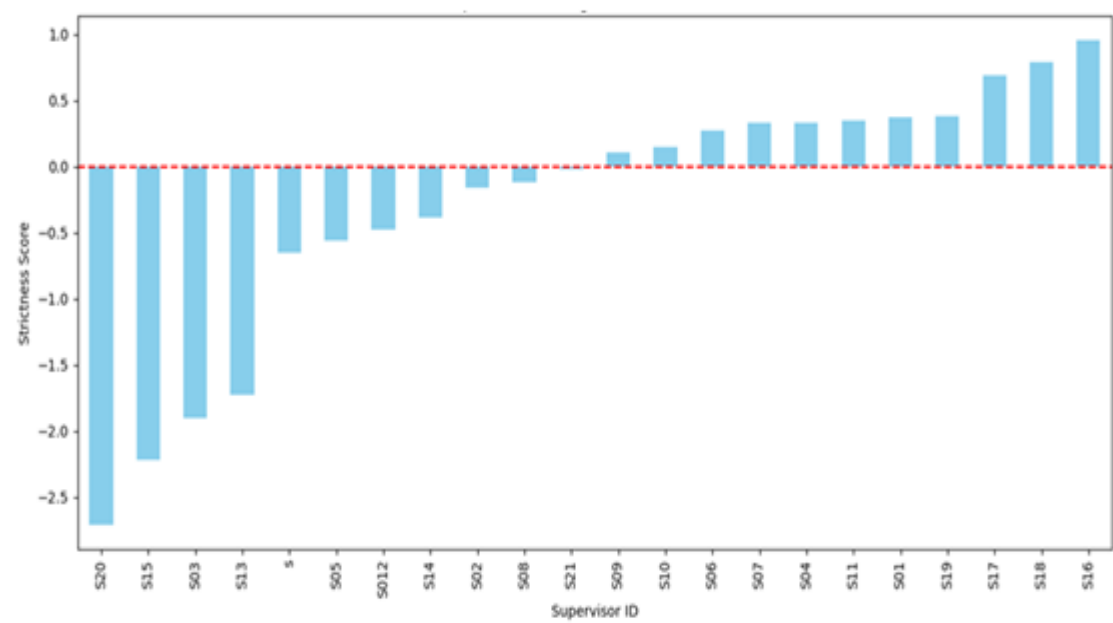


Figure 5: Supervisors ranked by standardised scoring strictness

Figure 5. Supervisors ranked by standardised scoring strictness (Z-score of mean residual); negative values indicate stricter-than-average scoring and positive values indicate more lenient-than-average scoring relative to the model benchmark.

Supervisor Scoring Pattern Analysis

Figure 6 shows the average prediction residual (actual - predicted) for each supervisor. Positive residuals indicate that supervisors awarded higher scores than the model benchmark (more lenient scoring), while negative residuals indicate that supervisors awarded lower scores than the benchmark (stricter scoring). The results reveal differences in scoring behaviour among supervisors. Supervisor S20 has the highest positive residual (+4.92), indicating a stronger tendency to award higher scores than the benchmark and suggesting possible

leniency. In contrast, Supervisor S012 has the largest negative residual (−2.69), indicating a tendency to assign lower scores than the benchmark, suggesting possible strictness. The remaining supervisors (S01, S19, S04, S10, S21, and S18) have small positive residuals close to zero, indicating only slight leniency and generally scoring close to the benchmark. Similarly, S06, S16, S05, and S02 have small negative residuals near zero, indicating slightly stricter scoring with minimal deviation from the model benchmark.

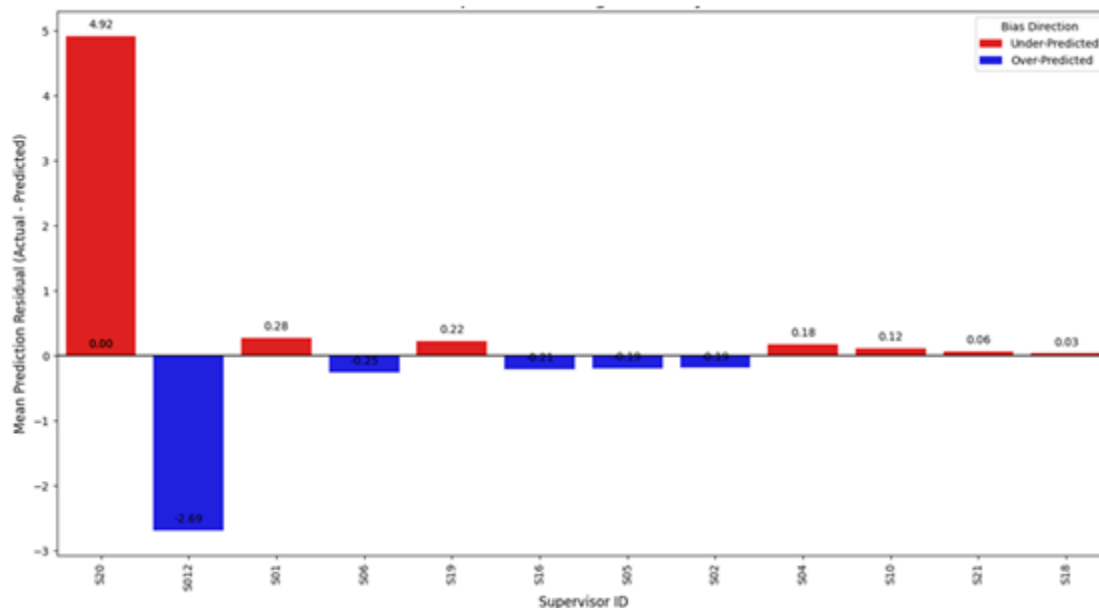


Figure 6: Mean prediction residual

Figure 6 shows the mean prediction residual (actual-predicted) per supervisor, ordered by magnitude; supervisors with the largest deviations (S20, S012) are highlighted as candidates for moderation review under the proposed ± 1.5 SD flagging rule described in the Methodology, pending formal confirmation once that threshold is applied to the full residual distribution. Most supervisors demonstrate scoring practices close to the model benchmark, with only minor deviations. However, S20 and S012 stand out as potential outliers, showing a benchmark deviation large enough to warrant moderation or further review. This pattern is currently established descriptively rather than through a formal inferential test (see Limitations); a Kruskal–Wallis test of residuals, grouped by supervisor, is recommended to confirm that this variation exceeds ordinary sampling noise before the finding is treated as conclusive. These findings support the study’s conclusion that predictive analytics can help identify supervisors whose

scoring patterns differ from the pool norm, thereby supporting more targeted, evidence-based review in Teaching Practice assessment.

Discussion

The findings support the potential of predictive modelling to surface irregularities in Teaching Practice assessment records for further human review. The strong fit obtained by the XGBoost model to the pooled supervisor data demonstrates that TP scores can be closely approximated by a learned benchmark derived from assessment subcomponents. The study is broadly consistent with earlier research indicating that machine learning techniques can support the monitoring of assessment reliability and fairness in education. The findings align with García and Romero (2015), who emphasized the capacity of educational data mining to uncover hidden patterns in assessment processes. Similarly, the strong performance of XGBoost is consistent with findings by Chen and Guestrin (2016), who

established the algorithm's suitability for structured predictive tasks; this study extends that evidence to a new domain (TP assessment) and, in Table 5, directly benchmarks XGBoost against simpler alternatives rather than assuming its suitability from prior literature alone.

Residual analysis revealed that although most supervisors maintained scoring patterns close to the model benchmark, a smaller subset displayed repeated deviations. Such deviations may reflect inconsistent interpretation of assessment rubrics, grading leniency, severity effects, or contextual influences during evaluation, though, as emphasised throughout, a residual alone cannot distinguish among these explanations.

Our finding that a small subset of supervisors (S20, S012) show persistently large residuals, while the majority cluster close to the benchmark, agrees in shape with Magowan's (2023) finding that pandemic-era teacher-assigned grades diverged from model-predicted grades unevenly rather than uniformly across the population in both cases, a predictive benchmark reveals a minority of assessors/contexts as the primary source of divergence rather than a generalized problem across the whole population. Our findings differ from Batruch et al. (2023) in scope, however: their scoping review documents *why* discrimination in grading occurs (implicit stereotyping, contextual pressure) at a conceptual level, whereas the present residual-based approach can only establish *that* a divergence exists and *where* to look, not *why* — a limitation that should temper how the S20/S012 findings are interpreted operationally, as flags for review rather than conclusions about supervisor conduct.

The overall moderation-support design adopted here flagging candidates for human review rather than issuing automated verdicts is also consistent with the most recent work in this space. Olaniyan et al. (2026) proposed a comparable explainable, fairness-aware framework for exam score verification that combines semantic modeling, rubric alignment, and peer-consistency analysis to flag potential under- or over-scoring for human review; although their framework was built for text-based essay scoring using transformer embeddings rather than structured numeric rubrics, the shared design principle using a learned benchmark to surface candidates for review, then relying on human oversight and interpretable evidence rather than automated correction reinforces that the approach taken in this study reflects current best practice in AI-assisted assessment moderation rather than an idiosyncratic choice. The dominance of Presentation and Teacher Personality in the feature importance analysis suggests that supervisors, in aggregate, prioritise observable teaching behaviour over other pedagogical dimensions. While these components are essential, overreliance on visible performance indicators may inadvertently marginalise equally important dimensions such as instructional planning and assessment strategy. The study also highlights the practical importance of integrating predictive analytics into TP moderation systems. Traditional moderation approaches in many Colleges of Education are time-consuming and often unable to systematically identify hidden inconsistencies. Predictive models provide a scalable mechanism for identifying scoring patterns that warrant closer review. The variation in Z-scores indicates that supervisors

do not assess students uniformly relative to the pooled benchmark. While most supervisors exhibit scoring behaviour close to the average, a few are noticeably stricter or more lenient. This variation is, at present, observed descriptively rather than confirmed through a formal inferential test; as noted in the Limitations, a Kruskal–Wallis test is recommended before this pattern is treated as more than a preliminary signal. These findings support the use of predictive analytics to identify supervisors whose scoring patterns may benefit from moderation review, to improve fairness, consistency, and reliability in Teaching Practice assessment. Importantly, the study does not advocate replacing human supervisors with automated systems. Rather, predictive modelling should function as a decision-support mechanism that enhances transparency, accountability, and evidence-based moderation.

The implications of this research extend beyond Nigeria. Globally, educational institutions are increasingly exploring algorithmic approaches to strengthen assessment integrity and fairness. However, researchers have also warned that AI systems themselves may introduce bias if poorly designed or trained on unrepresentative data (Idowu, 2024). Consequently, predictive moderation systems should be implemented alongside ethical safeguards, periodic auditing, and human oversight.

Conclusion

This study developed a machine learning framework for detecting potential scoring inconsistency in Teaching Practice assessment within Nigerian Colleges of Education. Using

XGBoost regression as a behavioural benchmark, the study revealed that machine learning can closely approximate pooled supervisor scoring patterns and help identify scoring deviations that warrant further review. The findings indicate that predictive analytics can support fairness, reliability, and transparency in educational assessment by enabling evidence-based, targeted moderation rather than blanket manual review. The study contributes to the growing application of educational data mining in teacher education and provides a scalable framework for institutional quality assurance. This study has several limitations that should be addressed before the framework is treated as validated. First, the predictive framework was evaluated using only the XGBoost regression algorithm; it was not empirically benchmarked against alternative models such as Linear Regression, Decision Tree, or Random Forest on this dataset, so its advantage over simpler approaches is supported by precedent in the literature (Chen & Guestrin, 2016; Ordoñez-Avila et al., 2023) rather than by a head-to-head comparison. Second, model performance was assessed using a single 80:20 train-test split rather than cross-validation, and only MAE and R^2 were reported; RMSE and repeated (e.g., 10-fold) cross-validation are recommended additions for a future revision. Third, the identified scoring differences among supervisors were based on residual analysis and descriptive comparison (Figures 4–6) rather than a formal inferential statistical test; a Kruskal–Wallis test (or one-way ANOVA, if residuals are shown to be approximately normal) is needed to confirm that the observed variation across supervisors exceeds ordinary sampling noise. Fourth, a direct check of

whether the recorded Total TP Score exactly equals the arithmetic sum of its seven components was not performed; such a check would clarify why the model's R^2 (0.89) is not closer to the near-perfect value that would otherwise be expected. Fifth, the study used Teaching Practice data from a single institutional department, which may limit the generalizability of the findings. Future studies should address each of these points, incorporate larger multi-institutional datasets, and explore explainable AI techniques such as SHAP to clarify individual-prediction-level reasoning; recent systematic evidence (Türkmen, 2024) indicates that XAI methods such as SHAP and LIME are increasingly adopted across educational research to strengthen transparency and stakeholder trust, and recent frameworks combining explainability with fairness auditing in assessment contexts specifically (Kesgin et al., 2025; Olaniyan et al., 2026) provide a direct template for this extension.

References

Acar Güvendir, M., Kılıç, A. F., Güvendir, E., & Kaçak, T. (2026). Bridging minds and machines: A comparative study of AI and human rater agreement and reliability in educational assessment. *Education and Information Technologies*, 31(11), 3781–3804. <https://doi.org/10.1007/s10639-026-13949-7>

Recommendations

Based on the findings, the following recommendations are made:

1. Colleges of Education should integrate predictive analytics into TP moderation systems to support evidence-based assessment review, treating flagged cases as triggers for human review rather than automatic findings.
2. Supervisors should receive periodic training on rubric interpretation and consistency-awareness practices.
3. Educational institutions should institutionalise regular auditing of TP scoring patterns to improve assessment consistency.
4. National quality assurance bodies such as the NCCE should encourage the adoption of AI-assisted moderation frameworks, implemented alongside human oversight and periodic auditing.
5. Future systems should incorporate explainable AI (e.g., SHAP) and dashboard visualisation tools to improve transparency and usability.

Batruch, A., Autin, F., Bataillard, F., & Butera, F. (2023). Discrimination in grading: A scoping review of studies on teachers' discrimination in school. *Studies in Educational Evaluation*, 78, 101284. <https://doi.org/10.1016/j.stueduc.2023.101284>

Burrows, R., Gurevych, I., & Hancke, A. (2019). Automated essay scoring: A survey of the state of the art. *Computers & Education*, 137, 103–121. <https://doi.org/10.1016/j.compedu.2019.04.012>

Chen, T., & Guestrin, C. (2016). XGBoost: A

- scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>
- Ekubo, E. A., & Esiefarienrhe, B. M. (2022). Using machine learning to predict low academic performance at a Nigerian university. *The African Journal of Information and Communication*, 2022(30), 1–33.
- Fafunwa, A. B. (1974). *History of education in Nigeria*. George Allen & Unwin.
- García, S., & Romero, C. (2015). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(6), 982–992. <https://doi.org/10.1109/TSMC.2014.2376162>
- Idowu, J. A. (2024). Debiasing education algorithms. *International Journal of Artificial Intelligence in Education*, 34, 1510–1540. <https://doi.org/10.1007/s40593-023-00389-4>
- Jibril, M. (2026). Machine learning for predicting students' academic performance in selected educational technology courses in Nigerian universities. *Journal of Educational Technology Development and Exchange*, 19(2), 20–37. <https://doi.org/10.18785/jetde.1902.02>
- Kesgin, K., Kiraz, S., Kosunalp, S., & Stoycheva, B. (2025). Beyond performance: Explaining and ensuring fairness in student academic performance prediction with machine learning. *Applied Sciences*, 15, 8409. <https://doi.org/10.3390/app15158409>
- Magowan, L. (2023). Centre assessment grades in 2020: A natural experiment for investigating bias in teacher judgements. *Journal of Computational Social Science*, 6(2), 609–653. <https://doi.org/10.1007/s42001-023-00206-x>
- National Commission for Colleges of Education. (2012). *Minimum standards for Nigeria Certificate in Education (NCE)*. NCCE.
- Nitko, A. J., & Brookhart, S. M. (2007). *Educational assessment of students* (6th ed.). Pearson.
- Oko, J., & Aghenta, J. (2014). Teacher supervisors and the validity of teaching practice scores in Nigeria. *International Journal of Educational Assessment*, 9(1), 66–74.
- Okoli, A., & Eze, J. U. (2021). Assessment of student-teachers during teaching practice: Challenges for quality assurance in teacher education. *Nigerian Journal of Education and Practice*, 8(3), 22–28.
- Olaniyan, J., Verkijika, S. F., & Obagbuwa, I. C. (2026). An explainable fairness-aware AI framework for exam score verification. *Frontiers in Artificial Intelligence*, 9, 1824849. <https://doi.org/10.3389/frai.2026.1824849>
- Ordoñez-Avila, R., Salgado Reyes, N., Meza, J., & Ventura, S. (2023). Data mining techniques for predicting teacher evaluation in higher education: A systematic literature review. *Heliyon*, 9(3), e13939. <https://doi.org/10.1016/j.heliyon.2023.e13939>
- Türkmen, G. (2024). The review of studies on explainable artificial intelligence in educational research. *Journal of Educational Computing Research*, 63(2), 277–310. <https://doi.org/10.1177/07356331241310915>
- UNESCO. (2021). *Artificial intelligence in education: Challenges and*

- opportunities for sustainable development*. UNESCO.
- Zhen, Y., & Zhu, X. (2024). An ensemble learning approach based on TabNet and machine learning models for cheating detection in educational tests. *Educational and Psychological Measurement*, 84(4), 780–809. <https://doi.org/10.1177/00131644231191298>
- Zhou, T., & Jiao, H. (2023). Exploration of the stacking ensemble machine learning algorithm for cheating detection in large-scale assessment. *Educational and Psychological Measurement*, 83(4), 831–854. <https://doi.org/10.1177/00131644221117193>