Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Improving Rare Cheating Detection in Educational Assessments Through Psychometric Feature Engineering and Synthetic Oversampling

Muhammad Ali Gunawan

guns12380@gmail.com

STIK Kendal https://orcid.org/0000-0002-0801-4939

Ari Setiawan

Universitas Sarjanawiyata Tamansiswa https://orcid.org/0000-0002-3967-0998

I Putu Suiraoka

Poltekkes Kemenkes Denpasar https://orcid.org/0000-0002-1125-1038

Hawa Husna Ab Ghani

Sultan Zainal Abidin University https://orcid.org/0009-0000-2483-143X

Amalia Fitri

Universitas Pekalongan https://orcid.org/0009-0007-6361-1369

Dewi Azizah

Universitas Pekalongan

Research Article

Keywords: cheating detection, educational assessment, person-fit statistics, psychometric feature engineering, rare-event detection, synthetic oversampling, machine learning

Posted Date: June 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10045621/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

The rapid expansion of digital assessment has intensified concerns regarding academic integrity and the detection of dishonest behaviors in technology-enhanced learning environments. Rare cheating detection remains methodologically challenging because verified misconduct labels are difficult to obtain ethically and operationally, while cheating cases represent a small minority of assessment records. This study developed and evaluated a psychometric-informed machine learning framework using an empirical-parameterized synthetic data design. Real assessment data were first used to derive empirical distributions of latent ability, response consistency, item response behavior, and response-time patterns. These parameters were then used to generate a synthetic assessment dataset comprising 1,200 examinee records and 20 items, while cheating-related aberrance parameters were specified using prior evidence from person-fit, aberrant responding, and assessment-security research. The framework integrated latent ability estimates, person-fit statistics, anomaly scores, and response-time indicators within Logistic Regression, Random Forest, and Extreme Gradient Boosting models, with Synthetic Minority Oversampling Technique applied to address class imbalance. Results showed that psychometric feature engineering provided the strongest contribution to rare cheating detection. Person-fit statistics emerged as the most influential predictor, followed by anomaly scores, whereas response-time indicators contributed less when used alone. Synthetic oversampling improved minority-class sensitivity, especially for ensemble models, but its contribution was smaller than that of psychometric feature engineering. These findings support the value of integrating empirical psychometric parameters, theory-informed feature engineering, and interpretable machine learning for scalable academic integrity monitoring.

1. Introduction

The rapid expansion of digital learning environments has fundamentally transformed educational assessment practices across schools, universities, and professional certification systems. Online examinations, computer-based testing, remote proctoring platforms, and learning management systems are increasingly employed to evaluate student achievement, monitor learning progress, and support large-scale educational decision-making (Abdulgader Alsharif, 2025; Shard et al., 2024). These technologies provide substantial benefits in terms of scalability, efficiency, automated scoring, and immediate feedback (Vetrivel et al., 2025). Educational institutions have consequently accelerated the adoption of technology-enhanced assessment systems to support flexible learning pathways and expanding student populations (Jopp, 2020; Khupe et al., 2026; Rudinger, 2020). Growing reliance on digital assessment has simultaneously intensified concerns regarding the integrity, validity, and trustworthiness of assessment outcomes (Akhtar & Perea-Borobio, 2026; Fawns & Schaepkens, 2022; Schulz, 2023).

Academic integrity represents a fundamental requirement of educational assessment because assessment results serve as the basis for certification, placement, progression, and accountability decisions (Ellis et al., 2020; Sotiriadou et al., 2020; Stoesz, 2024). Violations of academic integrity

threaten the credibility of these decisions and undermine confidence in educational systems. The migration of assessments from controlled physical environments to digital platforms has created new opportunities for dishonest behavior, including unauthorized collaboration, answer sharing, impersonation, and the use of external resources during examinations (Dawson, 2020; Garg & Goel, 2022). Educational institutions therefore face increasing pressure to develop reliable and scalable mechanisms capable of identifying suspicious response behavior while preserving fairness and transparency (Kulkarni et al., 2025; Pradhan & Shah, 2025).

The consequences of undetected cheating extend beyond individual assessment outcomes. Invalid scores distort estimates of student achievement, bias instructional decisions, and compromise institutional accountability systems (Beach, 2021; M. D. Wang, 2025). Effective cheating detection has therefore become an important component of contemporary educational technology ecosystems. Traditional approaches to cheating detection have largely relied on psychometric diagnostics derived from Item Response Theory (IRT), particularly person-fit statistics that evaluate the consistency between observed responses and model expectations (Gorney et al., 2024; Hong et al., 2020; Jones et al., 2023). Indicators such as *lz* have demonstrated considerable utility for identifying aberrant response behavior because they directly capture deviations from theoretically expected response patterns (Doval & Delicado, 2020; Qiu et al., 2024; Sideridis & Alghamdi, 2025).

Recent advances in educational data mining and artificial intelligence have expanded opportunities for automated cheating detection (Du et al., 2020; Okewu et al., 2021). Machine learning algorithms can process large volumes of assessment data and identify complex patterns that are difficult to detect using conventional rule-based procedures (Zhang, 2024). Random Forest, Gradient Boosting, Support Vector Machines, and related techniques have shown promising performance across various educational prediction tasks, including risk identification, performance forecasting, and learner analytics (Albreiki et al., 2021; Fahd et al., 2022; Naicker et al., 2020). Nevertheless, machine learning models remain highly dependent on the quality of input features. Many existing studies rely primarily on aggregate performance indicators such as total scores, percentages of correct responses, or demographic variables, despite the availability of richer psychometric information derived from measurement models (Wilson, 2023; Zesch et al., 2023).

An important limitation of the current literature is the insufficient integration between psychometric theory and machine learning practice. Psychometric approaches emphasize latent trait estimation, response consistency, and measurement validity (Christensen et al., 2020; Kyle et al., 2020), whereas machine learning approaches focus primarily on predictive optimization (Azevedo et al., 2024; Bian & Priyadarshi, 2024; Strielkowski et al., 2023). Consequently, educational assessment datasets often contain theoretically meaningful psychometric information that remains underutilized within predictive learning systems (Almalawi et al., 2024; Park et al., 2022). This separation between methodological traditions limits both the effectiveness and interpretability of cheating-detection models.

Psychometric feature engineering offers a promising mechanism for bridging this gap. Latent ability estimates summarize examinee proficiency while accounting for item characteristics and measurement uncertainty. Person-fit statistics provide complementary information regarding response-pattern consistency. Additional engineered indicators can further capture subtle irregularities that are difficult to observe through conventional score-based measures (Ozkat, 2024). Integrating these psychometric representations into machine learning pipelines enhances the detection of rare cheating behaviors.

A second challenge concerns class imbalance. In authentic educational settings, cheating typically represents a relatively small proportion of all assessment records. Classification models trained on highly imbalanced datasets frequently achieve satisfactory overall accuracy while failing to identify minority-class observations effectively(Milosevic & Ciric, 2022; Thabtah et al., 2020). Synthetic Minority Oversampling Technique (SMOTE) has emerged as one of the most widely adopted approaches for addressing this problem by generating synthetic minority observations during model training (Banerjee et al., 2020; Elreedy et al., 2024; Soltanzadeh & Hashemzadeh, 2021). Although oversampling can improve minority-class representation, its effectiveness depends heavily on the informational quality of the underlying feature space.

Simultaneously, growing concerns regarding transparency and accountability have increased interest in Explainable Artificial Intelligence (XAI) within educational applications (Chaushi et al., 2023; Kalasampath et al., 2025). Educational stakeholders increasingly require interpretable evidence explaining why a predictive model identifies a particular observation as suspicious. Feature-importance analyses and SHAP-based approaches have therefore become valuable tools for examining the mechanisms underlying machine learning decisions (Alomari & Andó, 2024; H. Wang et al., 2024).

This study develops and evaluates a framework that combines psychometric feature engineering and synthetic oversampling for rare cheating detection in educational assessments. The proposed framework integrates latent ability estimates, person-fit statistics, anomaly scores, and response-time indicators within machine learning classification models. Logistic Regression, Random Forest, and Extreme Gradient Boosting are subsequently evaluated to examine the relative contribution of psychometric representations and imbalance-aware learning strategies.

The study contributes to the literature in four important ways. First, it demonstrates how psychometric feature engineering can be integrated into machine learning pipelines for academic integrity monitoring. Second, it evaluates the relative predictive value of psychometric and behavioral indicators for rare cheating detection. Third, it examines the extent to which synthetic oversampling improves minority-class sensitivity under realistic class-imbalance conditions. Fourth, it provides explainable evidence regarding the relative importance of psychometric and behavioral features through machine learning interpretability analyses. Collectively, these contributions advance understanding of how psychometric theory and educational artificial intelligence can be combined to support trustworthy, interpretable, and scalable assessment-integrity systems.

2. Theoretical Background and Research Questions

2.1 Academic Integrity and Cheating Detection in Digital Assessments

The rapid expansion of digital assessment has increased concern regarding academic integrity and the reliability of assessment outcomes. Online examinations, computer-based testing, and remote assessment systems have created new opportunities for academic misconduct, including unauthorized collaboration, answer sharing, impersonation, and the use of external resources during examinations (Butler-Henderson & Crawford, 2020; Leong, 2026). These challenges have motivated growing interest in automated approaches capable of identifying suspicious behaviors using assessment data.

Traditional cheating-detection procedures frequently rely on human monitoring, rule-based systems, or post-examination investigations. Although such approaches remain useful, they often lack scalability and fail to detect subtle forms of misconduct in large-scale digital learning environments (Tiong et al., 2024). Educational researchers have therefore increasingly adopted data-driven techniques that leverage assessment records, behavioral traces, and psychometric indicators to identify anomalous response patterns (Akhuseyinoglu & Brusilovsky, 2022; Lin et al., 2024; Sajja et al., 2025).

Recent developments in educational data mining and artificial intelligence have further expanded opportunities for predictive cheating detection. Machine learning algorithms can identify complex nonlinear relationships and process large volumes of educational data efficiently. Nevertheless, model performance depends fundamentally on the quality and educational relevance of the available features. Consequently, feature construction has emerged as a central challenge in educational predictive analytics.

2.2 Person-Fit Theory and Psychometric Indicators

Psychometric theory provides one of the most established frameworks for detecting aberrant response behavior. Within Item Response Theory (IRT), examinee responses are expected to follow probabilistic patterns determined by latent ability and item characteristics. Departures from these expected patterns suggest unusual test-taking behaviors that warrant further investigation (Han & Kang, 2023; Pokropek & Khorramdel, 2025).

Person-fit statistics were developed to evaluate the consistency of individual response patterns relative to psychometric model expectations (Gao et al., 2025; Gorney & Wollack, 2023; Jones et al., 2023). These statistics quantify the extent to which observed responses conform to the behavior predicted by an IRT model and have been widely used to identify response irregularities associated with guessing, carelessness, item preknowledge, answer copying, and other forms of aberrant responding.

A substantial body of empirical research has demonstrated the value of person-fit statistics for assessment-integrity monitoring. Beck et al. (2019) reported that person-fit indicators provide reliable

evidence of aberrant response behavior across a variety of testing conditions. Similar findings have been observed in high-stakes examinations where response-pattern inconsistency frequently serves as an indicator of potential misconduct. More recent studies have further shown that aberrant response patterns can compromise measurement quality, distort psychometric estimates, and threaten the validity of score interpretations. Consequently, person-fit statistics remain among the most theoretically grounded tools available for identifying irregular testing behavior.

Among the various person-fit statistics proposed in the psychometric literature, the standardized *lz* statistic has received particular attention because it is firmly rooted in IRT theory, readily interpretable, and applicable across diverse assessment settings. The *lz* statistic evaluates the likelihood of an observed response pattern under a specified measurement model, with extreme values indicating substantial departures from expected response behavior. Previous studies have successfully employed *lz* to identify unusual response patterns associated with guessing, carelessness, item preknowledge, and potential cheating behavior (He et al., 2022; Sideridis & Alghamdi, 2025). These characteristics make *lz* a suitable psychometric indicator for feature engineering within predictive assessment models.

Despite their demonstrated effectiveness, person-fit statistics are commonly used as screening or diagnostic tools rather than as inputs for predictive machine-learning systems. As a result, theoretically meaningful psychometric evidence often remains disconnected from contemporary educational artificial intelligence approaches. Integrating person-fit information into predictive modeling therefore represents an important opportunity to combine measurement theory with advances in educational data analytics.

2.3 Psychometric Feature Engineering for Educational Artificial Intelligence

Machine learning performance is strongly influenced by the quality of input representations. Research in educational data mining has consistently shown that meaningful educational features frequently contribute more to predictive performance than increases in algorithmic complexity (Hilbert et al., 2021; Roy & Farid, 2024). Feature engineering has therefore become a central component of educational artificial intelligence.

Psychometric feature engineering extends this principle by transforming measurement-theoretic evidence into analytically useful inputs for predictive models. Latent ability estimates summarize examinee proficiency while accounting for item characteristics and measurement uncertainty. Person-fit statistics reflect the degree to which observed response patterns conform to psychometric model expectations. These measures provide theoretically grounded descriptions of assessment behavior that are not readily captured through conventional achievement scores. Educational assessment systems can also benefit from incorporating behavioral measures and assessment-derived indicators that reflect response processes and unusual testing patterns.

Recent scholarship has emphasized the value of integrating psychometric theory with machine learning. Edelsbrunner & Dablander (2019) argued that psychometric evidence offers conceptually meaningful

inputs capable of strengthening educational prediction models. Similar observations have been reported in adaptive assessment, learner analytics, and educational recommendation systems, where latent-trait measures contribute to both model performance and interpretability. Studies of aberrant responding further indicate that irregular testing behavior leaves detectable traces within assessment records, creating opportunities for feature construction that is informed by substantive measurement theory rather than purely data-driven selection procedures.

Cheating behavior rarely emerges through a single observable signal. Instead, it is often reflected in combinations of response inconsistencies, unusual timing patterns, and other departures from expected assessment behavior. Integrating psychometric measures with complementary behavioral and assessment-derived indicators is therefore expected to provide a more comprehensive representation of potential cheating behavior and improve the identification of rare irregular events. Evaluating the extent to which psychometric-informed feature engineering enhances predictive performance is consequently an important methodological objective in educational assessment analytics. Accordingly, the following research question was formulated:

RQ1

To what extent does psychometric feature engineering improve rare cheating detection performance in educational assessments?

2.4 Class Imbalance and Synthetic Oversampling

Rare-event classification remains a major challenge in educational analytics. In authentic assessment environments, cheating typically represents a relatively small proportion of all observations. Under such conditions, classification algorithms often become biased toward majority-class observations and exhibit poor sensitivity to minority-class events despite achieving high overall accuracy (Johnson & Khoshgoftaar, 2019; Sharma & Gosain, 2025).

Synthetic Minority Oversampling Technique (SMOTE) is among the most widely used approaches for addressing class imbalance. Rather than duplicating minority observations, SMOTE generates synthetic examples by interpolating between neighboring minority-class instances, thereby increasing class representation while preserving local data structure (Elreedy et al., 2024; H. Guan et al., 2021; Soltanzadeh & Hashemzadeh, 2021).

Extensive evidence supports the effectiveness of oversampling techniques in rare-event classification. Fernández et al. (2018) demonstrated that oversampling frequently improves minority-class detection across a variety of machine learning algorithms (de Morais & Vasconcelos, 2019; Zhu et al., 2022). Similar improvements have been reported in educational prediction studies involving dropout detection, student-risk identification, and academic-performance forecasting. Nevertheless, the effectiveness of oversampling depends strongly on the informational quality of the feature space. Informative predictors are often necessary before oversampling can generate meaningful performance improvements.

Severe class imbalance remains one of the primary challenges in cheating-detection research because minority-class observations are often underrepresented during model training. Synthetic oversampling techniques have been proposed to address this limitation by generating additional minority-class instances and improving class representation. Consequently, evaluating the extent to which synthetic oversampling improves the identification of rare cheating cases constitutes an important methodological question. Accordingly, the following research question was formulated:

RQ2

How does synthetic oversampling affect the detection of minority-class cheating cases across different machine-learning algorithms?

2.5 Explainable Artificial Intelligence and Feature Importance

The growing adoption of artificial intelligence in education has generated increasing concern regarding transparency, accountability, and interpretability (Mathew et al., 2025; Memarian & Doleck, 2023). Educational decisions often affect students, instructors, and institutions directly. Consequently, stakeholders frequently require explanations regarding how predictive models generate their classifications.

Explainable Artificial Intelligence (XAI) seeks to address this challenge by providing interpretable evidence regarding model behavior and feature contributions. Feature-importance analyses, SHAP values, and related approaches have become increasingly common within educational machine learning because they enable researchers to evaluate which variables drive prediction outcomes (Y. Guan et al., 2025; Takefuji, 2025; H. Wang et al., 2024).

From a psychometric perspective, explainability is particularly important because educational decisions should remain grounded in meaningful measurement constructs rather than opaque algorithmic mechanisms. If psychometric indicators genuinely capture cheating-related behaviors, they should contribute more strongly to model predictions than auxiliary behavioral indicators. Evaluating feature importance therefore provides both theoretical validation and practical insight into the mechanisms underlying cheating detection.

The literature reviewed above suggests that explainability analysis provides an important mechanism for evaluating the relative contribution of psychometric, behavioral, and assessment-derived indicators within cheating-detection systems. Examining feature importance therefore offers both theoretical insight and practical evidence regarding the factors that drive classification decisions. Accordingly, the following research question was formulated:

RQ3

Which psychometric, behavioral, and assessment-derived indicators contribute most strongly to rare cheating detection?

These research questions guide the evaluation of psychometric-informed feature engineering, synthetic oversampling, and explainable machine learning as complementary components of rare cheating detection in educational assessment systems.

3. Method

3.1 Research Design

This study employed an empirical-parameterized synthetic data design to evaluate the effectiveness of psychometric feature engineering and synthetic oversampling for rare cheating detection in educational assessments. This design was selected because verified cheating labels are difficult to obtain in operational educational settings, and experimentally inducing cheating behavior raises serious ethical and practical concerns. It is neither appropriate nor educationally defensible to require examinees to engage in dishonest behavior or to manipulate established assessment settings solely for research purposes. For this reason, the present study used real assessment data to derive empirical psychometric and behavioral parameters, which were then used to generate a controlled synthetic dataset that preserved the statistical structure of the original assessment environment.

The design consisted of two complementary components. The first component involved empirical calibration using real assessment data. This stage produced parameter estimates related to latent ability, response consistency, item response behavior, and response-time patterns. The second component involved synthetic data generation using the empirically derived parameters and cheating-related aberrance assumptions informed by prior research on person-fit statistics, aberrant responding, item preknowledge, response irregularity, and assessment-security violations. This approach allowed the study to preserve empirical realism while creating rare-event conditions required for evaluating cheating-detection models under class imbalance.

The proposed framework integrated psychometric indicators derived from Item Response Theory (IRT), behavioral response-time information, machine learning classification algorithms, Synthetic Minority Oversampling Technique (SMOTE), and Explainable Artificial Intelligence (XAI). The primary objective was to determine whether psychometric feature engineering improves rare cheating detection performance and to identify the relative contribution of psychometric and behavioral indicators within interpretable machine learning classification systems.

The analytical workflow consisted of six stages: (1) empirical parameter calibration, (2) synthetic assessment dataset generation, (3) psychometric feature extraction, (4) psychometric-informed feature engineering, (5) machine learning classification with and without synthetic oversampling, and (6) explainability analysis. This workflow provided a structured basis for evaluating the contribution of

feature representation, imbalance handling, and interpretability in rare-event academic integrity monitoring.

3.2 Empirical Parameter Calibration and Synthetic Dataset Construction

The dataset used in this study was generated through an empirical-parameterized procedure rather than a purely artificial simulation. Real educational assessment data were first used to estimate the empirical distributions and dependency structures needed for data generation. These empirical inputs included latent ability distributions, item response characteristics, response-pattern consistency, and response-time behavior. The purpose of this stage was to ensure that the generated dataset retained the statistical properties of authentic assessment data rather than relying on arbitrary assumptions.

The empirical calibration stage provided baseline parameters for normal assessment behavior. These parameters were used to model examinees whose response patterns were consistent with expected test-taking behavior. Cheating-related and aberrant response patterns were then introduced through parameter adjustments informed by prior studies on cheating detection, person-fit statistics, aberrant responding, response-time irregularity, and assessment-security research. This procedure allowed the generated data to reflect both the structure of real assessment behavior and theoretically plausible deviations associated with academic misconduct.

The final synthetic assessment dataset consisted of 1,200 examinee records and 20 assessment items. Four latent behavioral groups were represented: Honest, Mild Aberrant, Moderate Aberrant, and Severe Aberrant. Honest examinees represented empirically calibrated normal assessment behavior. The three aberrant groups represented increasing levels of response irregularity associated with potential cheating-related behavior. The final class distribution contained 1,080 honest examinees and 120 aberrant examinees, producing a 9:1 class imbalance ratio. This distribution was chosen to approximate rare-event conditions commonly encountered in educational integrity monitoring, where suspected cheating represents a small proportion of total assessment records.

This design provided a controlled analytical environment for comparing machine learning models while retaining empirical grounding. Because the minority-class labels were generated within the data-generation framework, the labels should be interpreted as simulated aberrant or cheating-related cases rather than independently adjudicated misconduct cases. The study therefore evaluates rare cheating detection under empirically grounded and controlled conditions, not as a final operational validation of misconduct detection in live assessment systems.

3.3 Psychometric Feature Extraction

Psychometric feature extraction was conducted to transform response data into theoretically meaningful indicators of examinee behavior. Three primary indicators were extracted: latent ability estimates, person-fit statistics, and response-time indicators. These features were selected because

they represent complementary aspects of assessment behavior: proficiency, response consistency, and behavioral timing.

Latent Ability Estimate

Latent ability estimates (θ) were used to represent examinee proficiency while accounting for differences in item difficulty and discrimination. Ability estimation was based on the two-parameter logistic Item Response Theory model:

$$P(X_{ij} = 1 \mid \theta_i, a_j, b_j) = \frac{1}{1 + \exp[-a_j(\theta_i - b_j)]}$$

where $P(X_{ij} = 1)$ denotes the probability that examinee i correctly answers item j, a_j represents the discrimination parameter of item j, b_j represents item difficulty, and θ_i represents the latent ability of examinee i. The latent ability estimate provided a psychometric representation of achievement that was separated from response-pattern irregularity.

Person-Fit Statistic

Response-pattern consistency was evaluated using the standardized person-fit statistic lz. Person-fit statistics quantify the degree to which observed item responses conform to expectations under an IRT model. The statistic was defined as:

$$lz_i = \frac{l_i - E(l_i)}{\sqrt{Var(l_i)}} \quad l_{zi} = \frac{l_i - E(l_i)}{\sqrt{Var(l_i)}}$$

Where

l_i = log-likelihood of examinee i's observed response pattern,

$E(l_i)$ = expected value, and

$Var(l_i)$ = variance under the fitted IRT model.

Lower values indicate stronger departures from expected response behavior and therefore provide evidence of potential aberrant responding.

Mean Response Time

Behavioral timing information was represented using mean response time. For each examinee, mean response time was calculated as:

$$\bar{RT}_i = \frac{1}{J} \sum_{j=1}^{J} RT_{ij}$$

where RT_{ij} denotes the response time of examinee i on item j, and J denotes the total number of assessment items. Response-time behavior was included because unusual timing patterns are frequently associated with item preknowledge, rapid guessing, external assistance, or other irregular test-taking behaviors.

3.4 Psychometric-Informed Feature Engineering

The feature engineering process was designed to represent multiple dimensions of assessment behavior relevant to rare cheating detection. The resulting feature set incorporated psychometric indicators, behavioral indicators, and assessment-derived anomaly indicators. Psychometric indicators were derived from IRT-based estimation and person-fit analysis, whereas behavioral information was represented through response-time measures.

The anomaly score was included as an assessment-derived indicator designed to capture broader irregularity beyond conventional psychometric and timing measures. Unlike latent ability and person-fit statistics, which are derived directly from IRT-based response modeling, the anomaly score represented a composite signal of abnormal assessment behavior. This score was generated using empirically calibrated distributional assumptions and group-specific aberrance severity parameters informed by previous cheating-detection and assessment-security studies. The anomaly score was specified as:

$$Anomaly_i = \max(0, Z_i + \delta_g)$$

where Z_i denotes the latent anomaly component generated from a multivariate distribution calibrated to the empirical assessment data, and δ_g denotes the group-specific aberrance severity adjustment for examinee group g. The non-negative constraint ensured that anomaly scores represented the degree of irregularity rather than a bidirectional deviation. The severity parameters were specified as follows:

Table 1
Group-Specific Aberrance Severity Adjustments Used in the Empirical-Parameterized Data-Generation Framework

Group	Severity Adjustment
Honest	0.0
Mild Aberrant	0.4
Moderate Aberrant	0.8
Severe Aberrant	1.5

These severity values were designed to represent increasing levels of response irregularity. Honest examinees were assigned no severity adjustment, whereas severe aberrant examinees received the largest adjustment. This structure allowed the generated data to contain graded forms of irregularity rather than treating all aberrant cases as homogeneous.

The engineered features were subsequently integrated into machine learning classification models. This approach enabled direct evaluation of whether theoretically grounded psychometric features improved rare cheating detection compared with isolated behavioral timing information.

3.5 Correlation Structure

The data-generation procedure was designed to reproduce theoretically plausible and empirically grounded relationships among psychometric and behavioral indicators. A predefined latent correlation structure was imposed among latent ability, person-fit statistics, anomaly scores, and response-time measures. The correlation matrix was specified based on empirical assessment data and previous studies on aberrant responding, person-fit analysis, and cheating-related behavior. The target correlation structure was specified as follows:

Table 2
Target Correlation Structure Used in the Empirical-Parameterized Data-Generation Framework

Variable	**θ**	**lz**	**Anomaly**	**RT**
θ	**1.00**	**0.20**	**−0.10**	**0.10**
lz	0.20	1.00	−0.40	0.15
Anomaly	−0.10	−0.40	1.00	−0.20
RT	0.10	0.15	−0.20	1.00

This structure reflected three assumptions. First, latent ability and person-fit statistics were allowed to correlate moderately because examinee proficiency is related to response-pattern consistency. Second, anomaly scores were specified to correlate negatively with person-fit statistics because greater irregularity is expected to coincide with lower response-pattern fit. Third, anomaly scores were specified to have weak relationships with latent ability and response time to ensure that the anomaly indicator captured irregularity rather than merely reproducing achievement or timing differences.

The correlation matrix was incorporated through Cholesky decomposition during data generation. This procedure ensured that generated examinee records preserved theoretically plausible dependencies among variables. The observed correlation matrix was then inspected in the Results section to evaluate whether the generated dataset reproduced the intended dependency structure.

3.6 Feature Configurations

Four feature configurations were evaluated to examine the incremental contribution of psychometric, behavioral, and assessment-derived information. This progressive feature construction strategy allowed direct comparison between psychometric indicators, behavioral timing indicators, and the full psychometric-informed feature engineering framework.

Table 3
Feature Configurations Evaluated in the Classification Models

Feature Set	Included Features	Purpose
FS1	θ, l_z	Psychometric baseline
FS2	Mean RT	Behavioral timing baseline
FS3	θ, l_z, Mean RT	Combined psychometric-behavioral configuration
FS4	θ, l_z, Mean RT, Anomaly Score	Full psychometric-informed feature engineering framework

The first configuration (FS1) assessed the predictive value of core psychometric indicators. The second configuration (FS2) tested the performance of response-time information as a behavioral baseline. The third configuration (FS3) evaluated whether combining psychometric and timing indicators improved rare cheating detection. The fourth configuration (FS4) represented the full proposed framework by integrating latent ability, person-fit statistics, response-time behavior, and anomaly scores.

3.7 Machine Learning Models

Three machine learning algorithms were evaluated: Logistic Regression, Random Forest, and Extreme Gradient Boosting. These models were selected because they represent distinct classification paradigms with different levels of complexity and interpretability. Logistic Regression served as a transparent baseline classifier. It estimated the probability of aberrant or cheating-related behavior using a linear decision boundary and provided interpretable model behavior. Random Forest represented a bagging-based ensemble approach that combines multiple decision trees through bootstrap aggregation and random feature selection. Extreme Gradient Boosting represented a boosting-based ensemble approach that sequentially constructs decision trees to capture nonlinear relationships and interaction structures.

The comparison among these models allowed the study to evaluate whether rare cheating detection performance was driven primarily by algorithmic complexity or by the quality of psychometric feature representation. This distinction was central to the study because the proposed framework emphasizes theory-informed feature engineering rather than reliance on algorithmic sophistication alone.

3.8 Training and Validation Procedure

The dataset was randomly partitioned into training and testing subsets using an 80:20 holdout strategy. The training set consisted of 960 examinee records, and the testing set consisted of 240 examinee records. Model development, threshold selection, and oversampling were conducted exclusively on the

training data. Final model performance was evaluated on the held-out testing data, which was not used during model training or oversampling.

Classification thresholds were selected using the training data to optimize minority-class detection while maintaining acceptable classification balance. The selected thresholds were subsequently applied to the held-out test set. This procedure prevented information leakage and ensured that performance metrics reflected generalization to unseen assessment records.

The use of a held-out test set was necessary because the study evaluated predictive performance under rare-event conditions. Metrics were therefore calculated only after models were applied to the unseen test data. This design provided a fair comparison across feature configurations, machine learning algorithms, and oversampling conditions.

3.9 Synthetic Oversampling

Synthetic Minority Oversampling Technique (SMOTE) was applied to the training data to address class imbalance. SMOTE generates synthetic minority-class observations by interpolating between neighboring minority-class instances within the feature space. This procedure increases minority-class representation without duplicating existing records.

Oversampling was performed only on the training dataset. The original testing dataset was not oversampled and remained unchanged throughout evaluation. This separation was essential to prevent information leakage and to ensure that test performance reflected model behavior under the original rare-event distribution.

SMOTE was used as a complementary strategy rather than as a replacement for feature engineering. The study specifically examined whether imbalance handling improved minority-class detection after psychometric feature engineering had been applied. This design allowed the relative contribution of feature representation and class rebalancing to be evaluated separately.

3.10 Evaluation Metrics

Classification performance was evaluated using multiple complementary metrics to capture overall discrimination, minority-class detection, and performance under class imbalance. ROC-AUC was used to evaluate overall discrimination across thresholds. PR-AUC was used because precision–recall performance is more informative than accuracy under rare-event conditions. Precision measured the proportion of predicted cheating-related cases that were correctly classified. Recall measured the proportion of true cheating-related cases that were successfully detected. F1-score summarized the balance between precision and recall. Matthews Correlation Coefficient provided an overall classification-quality index that accounts for all elements of the confusion matrix. Balanced Accuracy measured average performance across the majority and minority classes.

The use of multiple metrics was necessary because overall accuracy is misleading under severe class imbalance. A model can achieve high accuracy by classifying most cases as honest while failing to

detect rare cheating-related cases. For this reason, emphasis was placed on PR-AUC, Recall, F1-score, MCC, and Balanced Accuracy.

3.11 Explainable Artificial Intelligence Analysis

Explainability analyses were conducted to examine the mechanisms underlying model predictions and to evaluate the relative contribution of psychometric, behavioral, and assessment-derived indicators. Random Forest feature importance was evaluated using Mean Decrease Gini. Extreme Gradient Boosting feature importance was assessed using gain-based importance. SHAP analysis was used to estimate global feature contributions and examine the direction and magnitude of predictor effects.

The explainability analysis was intended to determine whether model predictions were driven by theoretically meaningful indicators rather than opaque algorithmic artifacts. Particular attention was given to the relative importance of latent ability, person-fit statistics, anomaly scores, and response-time indicators. If psychometric indicators genuinely capture cheating-related irregularities, they should contribute more strongly to model predictions than isolated timing-based features.

These analyses provided interpretive evidence for the proposed framework by linking predictive performance to substantively meaningful assessment constructs. This was especially important because academic integrity monitoring requires transparent and explainable evidence rather than black-box classification alone.

3.12 Ethical Rationale for the Empirical-Parameterized Design

The use of an empirical-parameterized synthetic data design was ethically justified by the nature of cheating-detection research. Directly inducing cheating behavior among examinees is not appropriate because it requires participants to engage in misconduct. Randomly manipulating authentic classroom or assessment settings also risks disrupting established educational processes and compromising fairness. Verified cheating labels are therefore difficult to obtain in routine educational assessments.

The present design addressed these constraints by separating empirical calibration from controlled synthetic data generation. Real assessment data were used to derive realistic psychometric and behavioral parameters, while prior research was used to specify plausible cheating-related aberrance patterns. This approach enabled rigorous model evaluation without exposing examinees to unethical experimental manipulation or requiring institutions to alter operational assessment conditions.

4. Results

4.1 Descriptive Statistics and Correlation Analysis

The simulation framework was designed to reproduce realistic educational assessment conditions while preserving meaningful variation in psychometric and behavioral characteristics. Considerable

heterogeneity was evident across all predictors. As reported in Table 4, latent ability estimates spanned a broad proficiency range, while person-fit statistics reflected substantial diversity in response consistency. Similar variability was observed for anomaly scores and response-time measures, suggesting that the simulated environment successfully captured multiple dimensions of assessment behavior rather than a narrow set of homogeneous observations.

Table 4
Descriptive Statistics of Psychometric and Behavioral Variables

Variable	Mean	SD	Minimum	Maximum
Latent Ability (θ)	-0.007	1.044	-3.230	3.032
Person-Fit Statistic (lz)	-0.141	1.038	-3.598	3.901
Anomaly Score	0.900	0.602	0.000	3.276
Mean Response Time	12.593	4.127	0.500	27.037

The distributions are noteworthy for another reason. None of the variables exhibited severe compression around a limited range of values, indicating that the feature space retained sufficient informational richness for classification purposes. From a modeling perspective, such diversity is advantageous because machine-learning algorithms benefit from predictors capable of separating minority and majority observations.

Relationships among the predictors provide additional insight into how these dimensions interact. Although several statistically meaningful associations emerged, the strongest relationship was observed between person-fit statistics and anomaly scores (Table 5).

Table 5
Correlation Matrix of Psychometric and Behavioral Features

Variable	θ(theta)	lz	Anomaly Score	Mean RT
θ(theta)	1	0.190	-0.075	0.138
Lz	0.190	1	-0.430	0.142
Anomaly Score	-0.075	-0.430	1	-0.203
Mean RT	0.138	0.142	-0.203	1

An important pattern emerges from the correlation structure. Anomaly scores exhibited a considerably stronger relationship with person-fit statistics than with latent ability estimates. Specifically, the moderate negative association between anomaly scores and person-fit statistics indicates that increasing behavioral irregularity tends to coincide with greater departures from model-expected

response patterns. Such a relationship is theoretically coherent because both indicators are designed to capture different dimensions of assessment aberrance.

As illustrated in Fig. 1, aberrance-related indicators occupy a noticeably different position from achievement-related measures within the broader correlation structure. By contrast, the near-zero association between anomaly scores and latent ability suggests that abnormal response behavior is largely independent of proficiency level. This distinction is important because it indicates that the proposed anomaly indicator is identifying behavioral irregularities rather than simply reflecting differences in achievement.

Correlation structure among psychometric and behavioral indicators used in the cheating-detection framework.

The visual representation reinforces the numerical findings by revealing a clustering of aberrance-related indicators while maintaining clear separation from ability and response-time measures. Such a configuration is advantageous because machine-learning models benefit from predictors that provide complementary rather than redundant information.

4.2 Baseline Performance of Psychometric and Behavioral Features

A central objective of this study was to determine whether psychometric feature engineering contributes meaningful predictive information beyond conventional psychometric and behavioral indicators. To address this question, four feature configurations were evaluated, ranging from a simple behavioral baseline to an integrated psychometric feature-engineering framework. The classification results reported in Table 6 reveal a clear performance gradient across feature configurations.

Table 6
Baseline Classification Performance Across Feature Configurations

Model	ROC_AUC	PRAUC	Precision	Recall	F1	MCC	BACC	Threshold
GLM_FS1	0.769	0.330	0.350	0.583	0.438	0.373	0.731	0.154
GLM_FS2	0.563	0.134	0.205	0.375	0.265	0.165	0.606	0.124
GLM_FS3	0.756	0.356	0.281	0.667	0.395	0.336	0.738	0.112
GLM_FS4	0.812	0.438	0.341	0.625	0.441	0.380	0.745	0.132

The behavioral baseline (FS2) produced the weakest performance across all evaluation metrics, indicating that response-time information provides only limited evidence for distinguishing between normal and aberrant assessment behavior. Although unusual timing patterns often accompany cheating, they appear insufficient to support reliable detection when used in isolation.

The inclusion of latent ability estimates and person-fit statistics (FS1) markedly improved discrimination and minority-class detection, suggesting that deviations from expected response behavior contain considerably richer information than timing-based measures. The strongest performance was achieved by the psychometric feature-engineering configuration (FS4). Integrating anomaly scores with traditional psychometric indicators generated the highest ROC-AUC, PR-AUC, F1-score, MCC, and balanced accuracy values. These improvements indicate that anomaly scores contribute information beyond conventional psychometric measures. Although person-fit statistics already capture departures from expected response patterns, anomaly scores appear to provide an additional layer of behavioral differentiation that enhances minority-class identification.

The findings suggest that feature construction exerts a stronger influence on cheating-detection performance than the inclusion of isolated behavioral indicators. The results therefore provide preliminary evidence supporting the value of psychometric feature engineering in educational artificial intelligence applications.

4.3 Comparison of Classification Algorithms

After identifying the most effective feature configuration, the next analysis examined whether predictive performance depended primarily on feature quality or algorithmic complexity. The comparative performance of Logistic Regression, Random Forest, and Extreme Gradient Boosting using the engineered feature set is presented in Table 7.

Table 7
Comparative Performance of Machine Learning Models

Model	ROC-AUC	PR-AUC	F1	MCC	BACC
GLM_FS4	0.812	0.438	0.441	0.380	0.745
RF_FS4	0.749	0.299	0.354	0.297	0.727
XGB_FS4	0.762	0.296	0.358	0.301	0.729

One of the most unexpected results was the superiority of Logistic Regression over both ensemble-learning approaches. Despite being the simplest classifier evaluated in this study, Logistic Regression consistently achieved the strongest overall performance across the primary evaluation metrics. Rather than benefiting from increasingly complex decision boundaries, the results suggest that predictive success depended primarily on the quality of the engineered psychometric features. Under such conditions, additional algorithmic sophistication contributed little incremental value to classification performance.

A plausible interpretation is that the engineered psychometric features already encoded much of the predictive structure required for cheating detection. When informative representations are available, increasingly complex decision boundaries contribute relatively little additional value because the primary signal has already been captured within the feature space. Random Forest and Extreme Gradient

Boosting nevertheless demonstrated competitive performance and maintained satisfactory discrimination capability. However, the relatively small performance differences among algorithms indicate that feature representation played a more influential role in cheating detection than model sophistication itself. This observation highlights the importance of theory-driven feature engineering within educational machine-learning systems.

4.4 Impact of Synthetic Oversampling on Rare Cheating Detection

Class imbalance represents one of the most persistent challenges in educational anomaly detection because cheating behaviors typically constitute only a small proportion of all assessment records. To evaluate whether minority-class representation could be improved, baseline models were compared with models trained using Synthetic Minority Oversampling Technique (SMOTE). Table 8 summarizes the resulting performance differences between the baseline and oversampled models.

Table 8
Impact of Synthetic Oversampling on Rare Cheating Detection

Model	ROC_AUC	PRAUC	Precision	Recall	F1	MCC	BACC
GLM_FS4_Baseline	0.812	0.438	0.341	0.625	0.441	0.380	0.745
RF_FS4_Baseline	0.749	0.299	0.236	0.708	0.354	0.297	0.727
XGB_FS4_Baseline	0.762	0.296	0.239	0.708	0.358	0.301	0.729
GLM_FS4_SMOTE	0.811	0.436	0.326	0.625	0.429	0.367	0.741
RF_FS4_SMOTE	0.761	0.291	0.276	0.667	0.390	0.331	0.736
XGB_FS4_SMOTE	0.761	0.355	0.194	0.875	0.318	0.285	0.736

The influence of oversampling varied across algorithms. Logistic Regression exhibited only minimal performance changes following SMOTE, indicating that the engineered psychometric features already provided substantial separation between cheating and non-cheating observations. The near-identical performance of the baseline and oversampled models suggests that class imbalance was not the primary limitation of the original framework.

More pronounced effects emerged for the ensemble-learning approaches. Random Forest showed modest improvements, whereas XGBoost responded much more strongly to oversampling. Recall increased from 0.708 to 0.875, indicating that a substantially larger proportion of cheating cases became detectable after minority-class representation was strengthened. The magnitude of this improvement highlights the value of oversampling when sensitivity to rare events is prioritized.

The improvement in sensitivity was accompanied by an expected trade-off. The considerable increase in recall observed for XGBoost was accompanied by reductions in precision and F1-score, indicating that improved sensitivity came at the cost of a higher false-positive rate. From an educational integrity

perspective, this trade-off is acceptable when the primary objective is early screening rather than final adjudication, because undetected cheating cases carry greater institutional consequences than a limited number of false-positive alerts.

The evidence indicates that psychometric feature engineering served as the primary driver of predictive performance, whereas SMOTE functioned as a complementary mechanism for increasing sensitivity to rare cheating events.

4.5 Feature Importance and Explainable Artificial Intelligence Analysis

Educational integrity monitoring systems require not only predictive accuracy but also transparency, interpretability, and theoretical justification. Equally important is understanding the factors that drive classification decisions. To address this requirement, explainability analyses were conducted using XGBoost, Random Forest, and SHAP-based interpretation techniques. The feature-importance structure derived from the XGBoost model is presented in Fig. 2.

Global feature contributions estimated from SHAP values in the Extreme Gradient Boosting model.

The XGBoost results indicate that predictive performance was concentrated within a relatively small subset of psychometric indicators. Person-fit statistics and anomaly scores accounted for the largest proportion of predictive information, whereas latent ability estimates and response-time measures contributed substantially less. Consistent with this pattern, Fig. 3 shows that Random Forest assigned the greatest importance to the same aberrance-related indicators, reinforcing the stability of the observed feature hierarchy across different learning algorithms.

Relative importance of psychometric and behavioral predictors in the Random Forest model.

The agreement between the two algorithms makes this pattern difficult to dismiss as a model-specific artifact. Similar rankings across independent learning mechanisms strengthen confidence that the dominance of aberrance-related indicators reflects a stable property of the data rather than a consequence of algorithm selection. The SHAP-based global importance visualization provides additional evidence regarding the relative contribution of each predictor to cheating classification.

Mean absolute SHAP values showing global feature contributions in the XGBoost model.

The concentration of influential SHAP values around person-fit statistics and anomaly scores demonstrates that these variables consistently drive prediction outcomes across a broad range of examinees. The global explanation further reveals that predictor influence remains relatively stable across observations, supporting the robustness of the proposed framework.

To summarize the explainability findings quantitatively, Table 9 reports the feature-importance rankings derived from both Random Forest and XGBoost models.

Table 9
Feature Importance Rankings Across Random Forest and XGBoost Models

Feature	RF MeanDecreaseGini	XGB Gain
lz	52.758	0.334
anomaly_score	49.895	0.333
theta	32.271	0.167
mean_rt	37.361	0.166

Across both algorithms, person-fit statistics and anomaly scores consistently emerged as the dominant predictors. The prominence of these indicators is theoretically meaningful because both variables directly reflect departures from expected assessment behavior. Although the relative ordering of latent ability and response-time measures differed slightly across algorithms, both remained substantially less influential than person-fit statistics and anomaly scores. This pattern underscores the dominant role of behavioral irregularities in classification decisions.

The explainability results converge on a clear conclusion: psychometric feature engineering exerted a stronger influence on cheating-detection performance than either algorithmic sophistication or synthetic oversampling. The dominance of person-fit statistics and anomaly scores across multiple analytical approaches provides a coherent empirical foundation for the discussion that follows.

5. Discussion

5.1 Psychometric Feature Engineering Improves Rare Cheating Detection Performance

The findings indicate that psychometric feature engineering substantially improved cheating-detection performance under empirically parameterized rare-event conditions. Feature configurations incorporating latent ability estimates, person-fit statistics, anomaly scores, and response-time indicators consistently outperformed simpler psychometric and behavioral baselines. More importantly, the strongest performance gains emerged after psychometric information was introduced into the feature space, indicating that predictive success depended more heavily on feature representation than on algorithmic sophistication.

One of the most unexpected findings was the superiority of Logistic Regression over both ensemble-learning approaches. Despite being the simplest classifier evaluated in the study, Logistic Regression consistently achieved the strongest overall performance across the primary evaluation metrics. An important implication is that the predictive signal embedded within the engineered features was sufficiently structured to be captured using relatively simple decision boundaries. Under such conditions, increasing model complexity offers limited benefit because the primary source of discrimination

originates from feature representation rather than nonlinear model architecture. This pattern shifts attention away from algorithm selection and toward the construction of educationally meaningful features as the principal determinant of cheating-detection performance.

A plausible explanation lies in the nature of cheating behavior itself. Cheating does not necessarily alter achievement outcomes in a consistent manner, but it often disrupts the coherence of expected response processes. Psychometric indicators such as person-fit statistics, anomaly scores, and latent ability estimates capture different manifestations of these disruptions. Person-fit statistics identify departures from model-expected response patterns, anomaly scores summarize multiple forms of behavioral irregularity, and latent ability estimates provide information regarding achievement levels. Together, these indicators create a multidimensional representation of assessment behavior that provides a more comprehensive characterization of potential cheating patterns than isolated performance scores or single-source behavioral measures. Such representations appear particularly valuable for identifying rare cheating events that remain undetected when relying exclusively on conventional assessment metrics.

Previous research in educational data mining has repeatedly shown that improvements in predictive performance are often driven more by theoretically informed feature construction than by the use of increasingly sophisticated algorithms. Studies by Guo et al. (2022), Wu et al. (2024), and Dai et al. (2024) reported similar advantages when educationally meaningful representations were used to model learner behaviors and assessment anomalies. The present study extends this body of work by demonstrating that psychometric representations derived from Item Response Theory remain highly effective even under rare-event conditions, where cheating cases constitute only a small fraction of assessment records.

Beyond improving predictive accuracy, the findings carry broader implications for the design of educational integrity systems. The results indicate that psychometric feature engineering can reduce dependence on increasingly sophisticated machine-learning architectures by embedding meaningful educational information directly within the feature space. This observation supports a growing movement toward theory-informed artificial intelligence, where advances in prediction originate not only from algorithmic innovation but also from stronger integration of domain knowledge. Assessment environments benefit particularly from this approach because predictive performance and interpretability can be improved simultaneously.

A contrasting perspective has been reported in several machine-learning studies where raw response data or clickstream information outperformed engineered features in large-scale digital learning environments. Such findings are typically observed in highly granular systems containing extensive behavioral logs and continuous interaction records. Educational assessments differ in important ways because available behavioral information is often considerably more limited. Under these conditions, psychometric feature engineering remains a highly effective mechanism for enhancing predictive performance while preserving transparency, interpretability, and theoretical coherence. The present findings therefore suggest that advances in educational artificial intelligence depend as much on

stronger integration of psychometric theory as on the development of increasingly sophisticated machine-learning algorithms.

5.2 The Limited but Meaningful Contribution of Synthetic Oversampling

Improvements attributable to synthetic oversampling were evident, particularly for minority-class detection, although their magnitude remained substantially smaller than the gains produced by psychometric feature engineering. Gains were most visible in recall-oriented metrics, particularly for XGBoost, where sensitivity to cheating cases increased substantially after oversampling. Nevertheless, the results showed that substantial predictive performance had already been achieved before oversampling was introduced. The modest effect of SMOTE reveals an important characteristic of rare cheating detection. Class imbalance is frequently treated as the primary obstacle in rare-event classification; however, the present findings suggest that the quality of educational representations exerts a stronger influence on performance than differences in class distribution. Models trained on psychometrically engineered features performed well even under imbalanced conditions, indicating that informative representations can partially mitigate the adverse effects of minority-class scarcity.

Research on imbalanced learning offers a useful context for understanding this pattern. Studies by Zhu et al. (2022), Alkhawaldeh et al. (2023), and Elyan et al. (2021) demonstrated that oversampling techniques improve minority-class recognition by increasing the visibility of rare observations during training. Similar benefits have been reported in educational applications involving dropout prediction, at-risk student identification, and other low-frequency events. The current findings confirm that cheating detection shares many of the same characteristics, particularly with respect to sensitivity improvements among minority cases.

At the same time, the results indicate that synthetic observations cannot compensate for weak feature representations. Oversampling generated only incremental improvements once psychometric feature engineering had already established strong separation between cheating and non-cheating observations. This distinction is important because it emphasizes that improvements in rare cheating detection depend not only on class representation but also on the development of theoretically meaningful educational features. Under such conditions, oversampling functions primarily as a complementary enhancement rather than the principal source of predictive performance.

Several studies have reported larger gains from SMOTE and related techniques, especially when minority-class observations contain limited informational content or when predictors provide weak class discrimination. The more moderate effects observed in the present study likely reflect the opposite situation. Psychometric feature engineering had already embedded substantial discriminatory information within the feature space before oversampling was applied. Consequently, the incremental value of SMOTE was naturally constrained by the strength of the original representation.

From a broader educational artificial intelligence perspective, these findings suggest that advances in rare-event detection should not focus exclusively on algorithmic remedies for class imbalance. Greater improvements emerge from combining robust educational representations with targeted imbalance-handling strategies. The evidence points to a central methodological lesson: rare-event detection improves most when imbalance handling is combined with theory-informed assessment features.

5.3 Person-Fit Statistics and Anomaly Scores as Dominant Predictors

Feature-importance analyses consistently identified person-fit statistics and anomaly scores as the most influential predictors of cheating behavior across Random Forest, XGBoost, and SHAP-based evaluations. The convergence of these independent explainability methods is particularly noteworthy because each algorithm relies on different learning mechanisms and variable-importance calculations. Despite these methodological differences, all approaches produced a remarkably similar ranking structure, indicating that cheating detection was driven primarily by indicators of response irregularity rather than by achievement level or response-time information.

The dominance of person-fit statistics is theoretically meaningful because cheating tends to disrupt the internal consistency of expected response processes before it substantially alters overall performance outcomes. Latent ability estimates summarize achievement levels, whereas person-fit statistics evaluate the degree to which observed response patterns conform to model-based expectations. Two examinees obtain similar scores while exhibiting very different response structures. In such situations, person-fit statistics provide evidence of response-pattern irregularities that cannot be inferred from achievement estimates. This distinction helps explain why person-fit indicators contributed more strongly to classification performance than latent ability measures throughout the analyses.

These findings are consistent with decades of psychometric research demonstrating the sensitivity of person-fit statistics to aberrant response behavior. Studies by Turner & Engelhard (2024), Gao et al. (2025), and Gorney & Wollack (2023) showed that person-fit indicators are effective for identifying unusual response patterns associated with guessing, item preknowledge, and various forms of test misconduct. The present study extends this literature by demonstrating that person-fit information remains highly valuable when incorporated into modern machine-learning frameworks designed for rare-event detection.

An equally important finding concerns the role of anomaly scores. Unlike traditional psychometric indicators that focus on a single dimension of assessment behavior, anomaly scores integrate multiple sources of irregularity into a unified representation. Such integration appears particularly valuable because cheating behavior rarely manifests through a single observable characteristic. Instead, suspicious responding often emerges through combinations of subtle deviations that individually appear unremarkable but become informative when considered collectively. The strong predictive contribution of anomaly scores suggests that composite representations of behavioral irregularity capture

dimensions of assessment behavior that are not fully represented by conventional psychometric statistics.

The relationship between person-fit statistics and anomaly scores is especially informative. Rather than competing sources of information, the two indicators appear to provide complementary evidence regarding assessment integrity. Person-fit statistics evaluate departures from expected response structures, whereas anomaly scores summarize broader manifestations of behavioral irregularity. Their simultaneous prominence across all explainability analyses suggests that successful cheating detection benefits from combining model-based psychometric evidence with integrated behavioral representations. This complementary relationship represents one of the most important contributions of the present study because it demonstrates how traditional measurement theory and contemporary machine-learning approaches can reinforce one another rather than operate as separate analytical paradigms.

Some behavioral analytics studies have reported response-time indicators as the strongest predictors of assessment misconduct. The current findings point toward a different interpretation. Response-time information contributed comparatively little once psychometric indicators and anomaly scores were available within the feature space. This discrepancy likely reflects differences in assessment environments and data availability. Highly granular digital learning systems often generate extensive behavioral traces that amplify the predictive value of timing information. Educational assessments typically provide more limited behavioral evidence, making psychometric indicators substantially more informative than response-time measures for distinguishing between normal and aberrant assessment behavior.

From an educational artificial intelligence perspective, these findings highlight the importance of psychometric-informed feature construction. Many predictive systems rely heavily on algorithmic sophistication while devoting comparatively less attention to the educational meaning of input variables. The present results suggest that explainable and trustworthy cheating-detection systems emerge not from increasingly complex algorithms, but from stronger integration of psychometric theory into predictive modeling frameworks. Person-fit statistics and anomaly scores therefore represent more than highly predictive variables; they illustrate how educational measurement principles can serve as a foundation for transparent and theoretically grounded artificial intelligence applications.

5.4 Implications for Educational Artificial Intelligence and Academic Integrity

Educational artificial intelligence systems designed to support assessment integrity face an increasingly difficult challenge: achieving high predictive performance without sacrificing interpretability. Contemporary educational AI research often emphasizes algorithmic innovation, model complexity, and predictive accuracy. The present results suggest a different perspective. Improvements in cheating detection were driven more strongly by psychometric feature representation than by increasingly sophisticated machine-learning architectures. This observation indicates that advances in educational AI

depend as much on the quality of educational knowledge embedded within predictive systems as on the complexity of the algorithms themselves.

Explainability emerges as a central issue when predictive systems are used to support academic misconduct investigations. Educational institutions are increasingly expected to justify automated decisions affecting learners, especially when those decisions involve allegations of academic misconduct. Black-box predictions achieve high classification accuracy while providing limited insight into why an examinee has been identified as suspicious. The current findings demonstrate that psychometrically grounded indicators such as person-fit statistics and anomaly scores can provide transparent evidence supporting model predictions. Consequently, explainability emerges not merely as a technical requirement but as an essential condition for responsible educational decision-making.

The results also contribute to a growing movement toward theory-informed artificial intelligence in education. Rather than treating educational data as generic prediction inputs, theory-informed approaches integrate established educational and psychometric principles directly into model development. The strong predictive value of person-fit statistics and anomaly scores suggests that educational measurement theory remains highly relevant in contemporary AI applications. Predictive systems informed by domain knowledge therefore achieve a more desirable balance between performance, interpretability, and theoretical coherence than systems relying exclusively on algorithmic optimization.

Implications extend beyond predictive modeling to institutional approaches for academic integrity monitoring. The findings suggest that artificial intelligence should function primarily as a decision-support mechanism rather than a replacement for professional judgment. Psychometric evidence can assist institutions in identifying potentially suspicious cases, but final determinations regarding misconduct should remain subject to human evaluation and contextual review. Such an approach reduces the risk of inappropriate sanctions while preserving the benefits of scalable integrity monitoring in large educational environments.

A more critical perspective highlights the potential risks associated with automated surveillance and algorithmic misconduct detection. Concerns regarding false positives, privacy, and procedural fairness remain highly relevant within educational settings. The present framework addresses these concerns by emphasizing interpretable indicators that can be independently examined and validated. Rather than producing opaque classifications, the model generates evidence-based signals grounded in established measurement theory. This distinction is particularly important because trust in educational AI depends not only on predictive performance but also on the transparency of the reasoning process underlying automated recommendations.

The findings position psychometrics and artificial intelligence as complementary foundations for the next generation of educational integrity systems. Traditional measurement theory offers principled methods for evaluating response validity, whereas modern machine-learning approaches provide scalable tools for identifying complex behavioral patterns. The present study demonstrates that these

traditions should be viewed as complementary rather than competing paradigms. Combining psychometric interpretability with computational predictive power provides a pathway toward integrity-monitoring frameworks that are simultaneously accurate, explainable, and educationally meaningful.

6. Conclusion

This study developed and evaluated a psychometric-informed machine learning framework for rare cheating detection using an empirical-parameterized synthetic data design grounded in real assessment parameters and prior cheating-detection evidence. The findings indicate that psychometric feature engineering provided the strongest contribution to detection performance under controlled rare-event conditions. Person-fit statistics and anomaly scores consistently emerged as the most influential predictors, while response-time indicators contributed less when used in isolation. Synthetic oversampling improved minority-class sensitivity, particularly for ensemble models, but its contribution remained smaller than the gains produced by theory-informed feature construction. These results support the central argument that effective academic integrity monitoring depends not only on algorithmic classification, but also on psychometrically meaningful representations of assessment behavior.

This study has several limitations. The final dataset was generated through an empirical-parameterized synthetic data procedure rather than obtained from operational assessment records containing independently adjudicated cheating cases. Although the data-generation process was grounded in real assessment parameters and prior evidence on aberrant responding, the generated labels should be interpreted as controlled cheating-related or aberrant cases, not as verified misconduct outcomes. The feature space was also limited to latent ability, person-fit statistics, anomaly scores, and mean response time. Additional process data, such as clickstream traces, item navigation behavior, answer revisions, keystroke dynamics, and multimodal proctoring indicators, were not incorporated. The analysis was further limited to three machine learning algorithms, which restricts conclusions regarding alternative anomaly-detection, cost-sensitive, Bayesian, graph-based, or deep learning approaches.

Future research should validate the proposed framework using operational assessment datasets containing independently confirmed cheating or misconduct cases. Further studies should examine whether combining multiple person-fit indicators, including (U3), (Zh), (Ht), (lz^*), and response-time-augmented person-fit statistics, improves detection performance beyond the current feature configuration. Future work should also evaluate the generalizability of psychometric-informed cheating detection across assessment formats, grade levels, subject areas, item types, and institutional settings. Because academic integrity decisions carry serious consequences, future implementation should position the proposed framework as a screening and decision-support tool that supports human review, procedural fairness, and transparent evidence-based investigation rather than as an autonomous misconduct adjudication system.

Declarations

Competing Interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethical Approval

Ethical approval was not required because the present analysis used an empirical-parameterized synthetic data design and did not involve the collection of identifiable human-subject data.

Informed Consent

Informed consent was not required for the present synthetic-data analysis because no identifiable human participants were directly involved.

Data Availability Statement

The empirical-parameterized data-generation procedures, scripts, and derived synthetic datasets are available from the corresponding author upon reasonable request.

References

1. Abdulgader Alsharif (2025) Artificial Intelligence and the Future of Assessment: Opportunities for Scalable, Fair, and Real-Time Evaluation. *Libyan Journal of Educational Research and E-Learning (LJERE)*, 42–52. https://doi.org/10.65417/ljere.v1i1.5
2. Akhtar Z, Perea-Borobio E (2026) Navigating the digital shift: investigating the viability of digital assessments for electrical and electronic engineering students. Eur J Eng Educ 51(3):714–733. https://doi.org/10.1080/03043797.2025.2523549
3. Akhuseyinoglu K, Brusilovsky P (2022) Exploring Behavioral Patterns for Data-Driven Modeling of Learners' Individual Differences. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.807320
4. Albreiki B, Zaki N, Alashwal H (2021) A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques. Educ Sci 11(9):552. https://doi.org/10.3390/educsci11090552
5. Alkhawaldeh IM, Albalkhi I, Naswhan AJ (2023) Challenges and limitations of synthetic minority oversampling techniques in machine learning. World J Methodol 13(5):373–378. https://doi.org/10.5662/wjm.v13.i5.373

6. Almalawi A, Soh B, Li A, Samra H (2024) Predictive Models for Educational Purposes: A Systematic Review. Big Data Cogn Comput 8(12):187. https://doi.org/10.3390/bdcc8120187
7. Alomari Y, Andó M (2024) SHAP-based insights for aerospace PHM: Temporal feature importance, dependencies, robustness, and interaction analysis. Results Eng 21:101834. https://doi.org/10.1016/j.rineng.2024.101834
8. Azevedo BF, Rocha AMAC, Pereira AI (2024) Hybrid approaches to optimization and machine learning methods: a systematic literature review. Mach Learn 113(7):4055–4097. https://doi.org/10.1007/s10994-023-06467-x
9. Banerjee A, Bhattacharjee M, Ghosh K, Chatterjee S (2020) Synthetic minority oversampling in addressing imbalanced sarcasm detection in social media. Multimedia Tools Appl 79(47–48):35995–36031. https://doi.org/10.1007/s11042-020-09138-4
10. Beach JM (2021) Can we measure what matters most? Why educational accountability metrics lower student learning and demoralize teachers. Bloomsbury Publishing PLC
11. Beck MF, Albano AD, Smith WM (2019) Person-Fit as an Index of Inattentive Responding: A Comparison of Methods Using Polytomous Survey Data. Appl Psychol Meas 43(5):374–387. https://doi.org/10.1177/0146621618798666
12. Bian K, Priyadarshi R (2024) Machine Learning Optimization Techniques: A Survey, Classification, Challenges, and Future Research Issues. Arch Comput Methods Eng. https://doi.org/10.1007/s11831-024-10110-w
13. Butler-Henderson K, Crawford J (2020) A systematic review of online examinations: A pedagogical innovation for scalable authentication and integrity. Comput Educ 159:104024. https://doi.org/10.1016/j.compedu.2020.104024
14. Chaushi BA, Selimi B, Chaushi A, Apostolova M (2023) *Explainable Artificial Intelligence in Education: A Comprehensive Review* (pp. 48–71). https://doi.org/10.1007/978-3-031-44067-0_3
15. Christensen AP, Golino H, Silvia PJ (2020) A Psychometric Network Perspective on the Validity and Validation of Personality Trait Questionnaires. Eur J Pers 34(6):1095–1108. https://doi.org/10.1002/per.2265
16. Dai C-P, Ke F, Pan Y, Moon J, Liu Z (2024) Effects of Artificial Intelligence-Powered Virtual Agents on Learning Outcomes in Computer-Based Simulations: A Meta-Analysis. Educational Psychol Rev 36(1):31. https://doi.org/10.1007/s10648-024-09855-4
17. Dawson P (2020) Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education. Routledge
18. de Morais RFAB, Vasconcelos GC (2019) Boosting the performance of over-sampling algorithms through under-sampling the minority class. Neurocomputing 343:3–18. https://doi.org/10.1016/j.neucom.2018.04.088
19. Doval E, Delicado P (2020) Identifying and Classifying Aberrant Response Patterns Through Functional Data Analysis. J Educational Behav Stat 45(6):719–749. https://doi.org/10.3102/1076998620911941

20. Du X, Yang J, Hung J-L, Shelton B (2020) Educational data mining: a systematic review of research and emerging trends. Inform Discovery Delivery 48(4):225–236. https://doi.org/10.1108/IDD-09-2019-0070
21. Edelsbrunner PA, Dablander F (2019) The Psychometric Modeling of Scientific Reasoning: a Review and Recommendations for Future Avenues. Educational Psychol Rev 31(1):1–34. https://doi.org/10.1007/s10648-018-9455-5
22. Ellis C, van Haeringen K, Harper R, Bretag T, Zucker I, McBride S, Rozenberg P, Newton P, Saddiqui S (2020) Does authentic assessment assure academic integrity? Evidence from contract cheating data. High Educ Res Dev 39(3):454–469. https://doi.org/10.1080/07294360.2019.1680956
23. Elreedy D, Atiya AF, Kamalov F (2024) A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Mach Learn 113(7):4903–4923. https://doi.org/10.1007/s10994-022-06296-4
24. Elyan E, Moreno-Garcia CF, Jayne C (2021) CDSMOTE: class decomposition and synthetic minority class oversampling technique for imbalanced-data classification. Neural Comput Appl 33(7):2839–2851. https://doi.org/10.1007/s00521-020-05130-z
25. Fahd K, Venkatraman S, Miah SJ, Ahmed K (2022) Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. Educ Inform Technol 27(3):3743–3775. https://doi.org/10.1007/s10639-021-10741-7
26. Fawns T, Schaepkens S (2022) A Matter of Trust: Online Proctored Exams and the Integration of Technologies of Assessment in Medical Education. Teach Learn Med 34(4):444–453. https://doi.org/10.1080/10401334.2022.2048832
27. Gao X, Hou M, Wang F, Zhou J (2025) A new person-fit statistic for the detection of aberrant responses in polytomous cognitive diagnostic models. Behav Res Methods 57(5):138. https://doi.org/10.3758/s13428-025-02659-6
28. Garg M, Goel A (2022) A systematic literature review on online assessment security: Current challenges and integrity strategies. Computers Secur 113:102544. https://doi.org/10.1016/j.cose.2021.102544
29. Gorney K, Sinharay S, Liu X (2024) Using item scores and response times in person-fit assessment. Br J Math Stat Psychol 77(1):151–168. https://doi.org/10.1111/bmsp.12320
30. Gorney K, Wollack JA (2023) Using Item Scores and Distractors in Person-Fit Assessment. J Educ Meas 60(1):3–27. https://doi.org/10.1111/jedm.12345
31. Guan H, Zhang Y, Xian M, Cheng HD, Tang X (2021) SMOTE-WENN: Solving class imbalance and small sample problems by oversampling and distance scaling. Appl Intell 51(3):1394–1409. https://doi.org/10.1007/s10489-020-01852-8
32. Guan Y, Wang F, Song S (2025) Interpretable machine learning for academic performance prediction: A SHAP-based analysis of key influencing factors. Innovations Educ Teach Int 1–20. https://doi.org/10.1080/14703297.2025.2532050

33. Guo T, Bai X, Tian X, Firmin S, Xia F (2022) Educational Anomaly Analytics: Features, Methods, and Challenges. *Frontiers in Big Data*, *4*. https://doi.org/10.3389/fdata.2021.811840
34. Han S, Kang H (2023) Online Monitoring of Test-Taking Behavior Based on Item Responses and Response Times. J Educ Meas 60(4):651–675. https://doi.org/10.1111/jedm.12367
35. He Q, Meadows M, Black B (2022) An introduction to statistical techniques used for detecting anomaly in test results. Res Papers Educ 37(1):115–133. https://doi.org/10.1080/02671522.2020.1812108
36. Hilbert S, Coors S, Kraus E, Bischl B, Lindl A, Frei M, Wild J, Krauss S, Goretzko D, Stachl C (2021) Machine learning for the educational sciences. Rev Educ 9(3). https://doi.org/10.1002/rev3.3310
37. Hong SE, Monroe S, Falk CF (2020) Performance of Person-Fit Statistics Under Model Misspecification. J Educ Meas 57(3):423–442. https://doi.org/10.1111/jedm.12207
38. Johnson JM, Khoshgoftaar TM (2019) Survey on deep learning with class imbalance. J Big Data 6(1):27. https://doi.org/10.1186/s40537-019-0192-5
39. Jones EA, Wind SA, Tsai C-L, Ge Y (2023) Comparing Person-Fit and Traditional Indices Across Careless Response Patterns in Surveys. Appl Psychol Meas 47(5–6):365–385. https://doi.org/10.1177/01466216231194358
40. Jopp R (2020) A case study of a technology enhanced learning initiative that supports authentic assessment. Teach High Educ 25(8):942–958. https://doi.org/10.1080/13562517.2019.1613637
41. Kalasampath K, Spoorthi KN, Sajeev S, Kuppa SS, Ajay K, Maruthamuthu A (2025) A Literature Review on Applications of Explainable Artificial Intelligence (XAI). IEEE Access 13:41111–41140. https://doi.org/10.1109/ACCESS.2025.3546681
42. Khupe D, Zlotnikova I, Hlomani H (2026) Advancing Technology-Enhanced Learning: A Systematic Literature Review of Innovative Frameworks and Emerging Technologies for Education 4.0. J Teach Learn 20(1). https://doi.org/10.22329/jtl.v19i3.9455
43. Kulkarni MD, Awate A, Shahade M, Nandwalkar B (2025) ExamGuard: Smart contracts for secure online test. Inform Syst 128:102485. https://doi.org/10.1016/j.is.2024.102485
44. Kyle G, Landon A, Vaske J, Wallen K (2020) Tools for assessing the psychometric adequacy of latent variables in conservation research. Conserv Biol 34(6):1353–1363. https://doi.org/10.1111/cobi.13625
45. Leong WY (2026) *E-Exams and Academic Integrity: Combating Cheating with Advanced Proctoring Solutions* (pp. 326–338). https://doi.org/10.1007/978-981-95-2011-4_31
46. Lin L, Zhou D, Wang J, Wang Y (2024) A Systematic Review of Big Data Driven Education Evaluation. Sage Open 14(2). https://doi.org/10.1177/21582440241242180
47. Mathew DE, Ebem DU, Ikegwu AC, Ukeoma PE, Dibiaezue NF (2025) Recent Emerging Techniques in Explainable Artificial Intelligence to Enhance the Interpretable and Understanding of AI Models for Human. Neural Process Lett 57(1):16. https://doi.org/10.1007/s11063-025-11732-2

48. Memarian B, Doleck T (2023) Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review. Computers Education: Artif Intell 5:100152. https://doi.org/10.1016/j.caeai.2023.100152
49. Milosevic MS, Ciric VM (2022) Extreme minority class detection in imbalanced data for network intrusion. Computers Secur 123:102940. https://doi.org/10.1016/j.cose.2022.102940
50. Naicker N, Adeliyi T, Wing J (2020) Linear Support Vector Machines for Prediction of Student Performance in School-Based Education. *Mathematical Problems in Engineering*, *2020*, 1–7. https://doi.org/10.1155/2020/4761468
51. Okewu E, Adewole P, Misra S, Maskeliunas R, Damasevicius R (2021) Artificial Neural Networks for Educational Data Mining in Higher Education: A Systematic Literature Review. Appl Artif Intell 35(13):983–1021. https://doi.org/10.1080/08839514.2021.1922847
52. Ozkat EC (2024) Vibration data-driven anomaly detection in UAVs: A deep learning approach. Eng Sci Technol Int J 54:101702. https://doi.org/10.1016/j.jestch.2024.101702
53. Park JY, Dedja K, Pliakos K, Kim J, Joo S, Cornillie F, Vens C, Van den Noortgate W (2022) Comparing the prediction performance of item response theory and machine learning methods on item responses for educational assessments. Behav Res Methods 55(4):2109–2124. https://doi.org/10.3758/s13428-022-01910-8
54. Pokropek A, Khorramdel L (2025) *Analyzing Test-Taking Behaviors Through Process Data Using the IRT Explanatory Model for Guessing* (pp. 559–584). https://doi.org/10.1007/978-3-031-90951-1_23
55. Pradhan V, Shah P (2025) A Comprehensive Review and Analysis of Blockchain Driven Identity Management Solutions for Enhancing Security and Transparency in Education. SN Comput Sci 6(8):924. https://doi.org/10.1007/s42979-025-04477-1
56. Qiu X, Huang S-Y, Wang W-C, Wang Y-G (2024) An Iterative Scale Purification Procedure on l_z for the Detection of Aberrant Responses. Multivar Behav Res 59(1):62–77. https://doi.org/10.1080/00273171.2023.2211564
57. Roy K, Farid D, Md (2024) An Adaptive Feature Selection Algorithm for Student Performance Prediction. IEEE Access 12:75577–75598. https://doi.org/10.1109/ACCESS.2024.3406252
58. Rudinger BJ (2020) *Technology-Enhanced Learning* (pp. 876–884). https://doi.org/10.1007/978-3-319-95870-5_62
59. Sajja R, Sermet Y, Cwiertny D, Demir I (2025) Integrating AI and Learning Analytics for Data-Driven Pedagogical Decisions and Personalized Interventions in Education. *Technology, Knowledge and Learning*. https://doi.org/10.1007/s10758-025-09897-9
60. Schulz A (2023) *Trust and Cheating: From Assessment Literacy to a Literacy of Practice for Digital Remote Assessments* (pp. 129–152). https://doi.org/10.1007/978-3-658-42253-0_7
61. Shard, Kumar D, Koul S (2024) Investigating students' perception about LMS-based online examination practices. Int J Inform Learn Technol 41(1):36–60. https://doi.org/10.1108/IJILT-12-2022-0219

62. Sharma S, Gosain A (2025) Addressing class imbalance in remote sensing using deep learning approaches: a systematic literature review. Evol Intel 18(1):23. https://doi.org/10.1007/s12065-024-01012-8
63. Sideridis G, Alghamdi MH (2025) Aberrant Responding in Hypothesis Testing: A Threat to Validity or Source of Insight? Behav Sci 15(3):319. https://doi.org/10.3390/bs15030319
64. Soltanzadeh P, Hashemzadeh M (2021) RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem. Inf Sci 542:92–111. https://doi.org/10.1016/j.ins.2020.07.014
65. Sotiriadou P, Logan D, Daly A, Guest R (2020) The role of authentic assessment to preserve academic integrity and promote skill development and employability. Stud High Educ 45(11):2132–2148. https://doi.org/10.1080/03075079.2019.1582015
66. Stoesz BM (2024) *Academic Integrity Through Ethical Teaching and Assessment: Overview and Current Trends* (pp. 203–216). https://doi.org/10.1007/978-3-031-54144-5_89
67. Strielkowski W, Vlasov A, Selivanov K, Muraviev K, Shakhnov V (2023) Rev Energies 16(10):4025. https://doi.org/10.3390/en16104025. Prospects and Challenges of the Machine Learning and Data-Driven Methods for the Predictive Analysis of Power Systems
68. Takefuji Y (2025) Beyond XGBoost and SHAP: Unveiling true feature importance. J Hazard Mater 488:137382. https://doi.org/10.1016/j.jhazmat.2025.137382
69. Thabtah F, Hammoud S, Kamalov F, Gonsalves A (2020) Data imbalance in classification: Experimental evaluation. Inf Sci 513:429–441. https://doi.org/10.1016/j.ins.2019.11.004
70. Tiong LCO, Lee Y, Lim KL, Lee HJ (2024) Advancing Online Assessment Integrity: Integrated Misconduct Detection via Internet Protocol Analysis and Behavioral Classification. IEEE Access 12:106056–106069. https://doi.org/10.1109/ACCESS.2024.3434608
71. Turner KT, Engelhard G (2024) Using Functional Clustering to Diagnose Person Misfit. J Experimental Educ 92(2):377–397. https://doi.org/10.1080/00220973.2022.2161088
72. Vetrivel SC, Arun VP, Ambikapathi R, Saravanan TP (2025) Automated grading systems: Enhancing efficiency and consistency in student assessments. Adopting Artificial Intelligence Tools in Higher Education. CRC, pp 41–61
73. Wang H, Liang Q, Hancock JT, Khoshgoftaar TM (2024) Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. J Big Data 11(1):44. https://doi.org/10.1186/s40537-024-00905-w
74. Wang MD (2025) Data quality disparities in large-scale assessments: insufficient effort responding across student groups, schools, and cultures. Large-Scale Assessments Educ 13(1):21. https://doi.org/10.1186/s40536-025-00260-z
75. Wilson M (2023) Constructing Measures. Routledge. https://doi.org/10.4324/9781003286929
76. Wu X, Chen C, Quan L (2024) Visual analysis and interactive interface design of students' abnormal behavior introducing clustering algorithm. Technol Health Care 32(6):4947–4963. https://doi.org/10.3233/THC-232054

77. Zesch T, Horbach A, Zehner F (2023) To Score or Not to Score: Factors Influencing Performance and Feasibility of Automatic Content Scoring of Text Responses. Educational Measurement: Issues Pract 42(1):44–58. https://doi.org/10.1111/emip.12544
78. Zhang J (2024) Evaluating Machine Learning Approaches for Sensitive Data Identification: A Comparative Study of NLP and Rule-Based Methods. J Adv Comput Syst 4(7):26–38. https://doi.org/10.69987/JACS.2024.40703
79. Zhu T, Liu X, Zhu E (2022) Oversampling with Reliably Expanding Minority Class Regions for Imbalanced Data Learning. IEEE Trans Knowl Data Eng 1–1. https://doi.org/10.1109/TKDE.2022.3171706

Figures

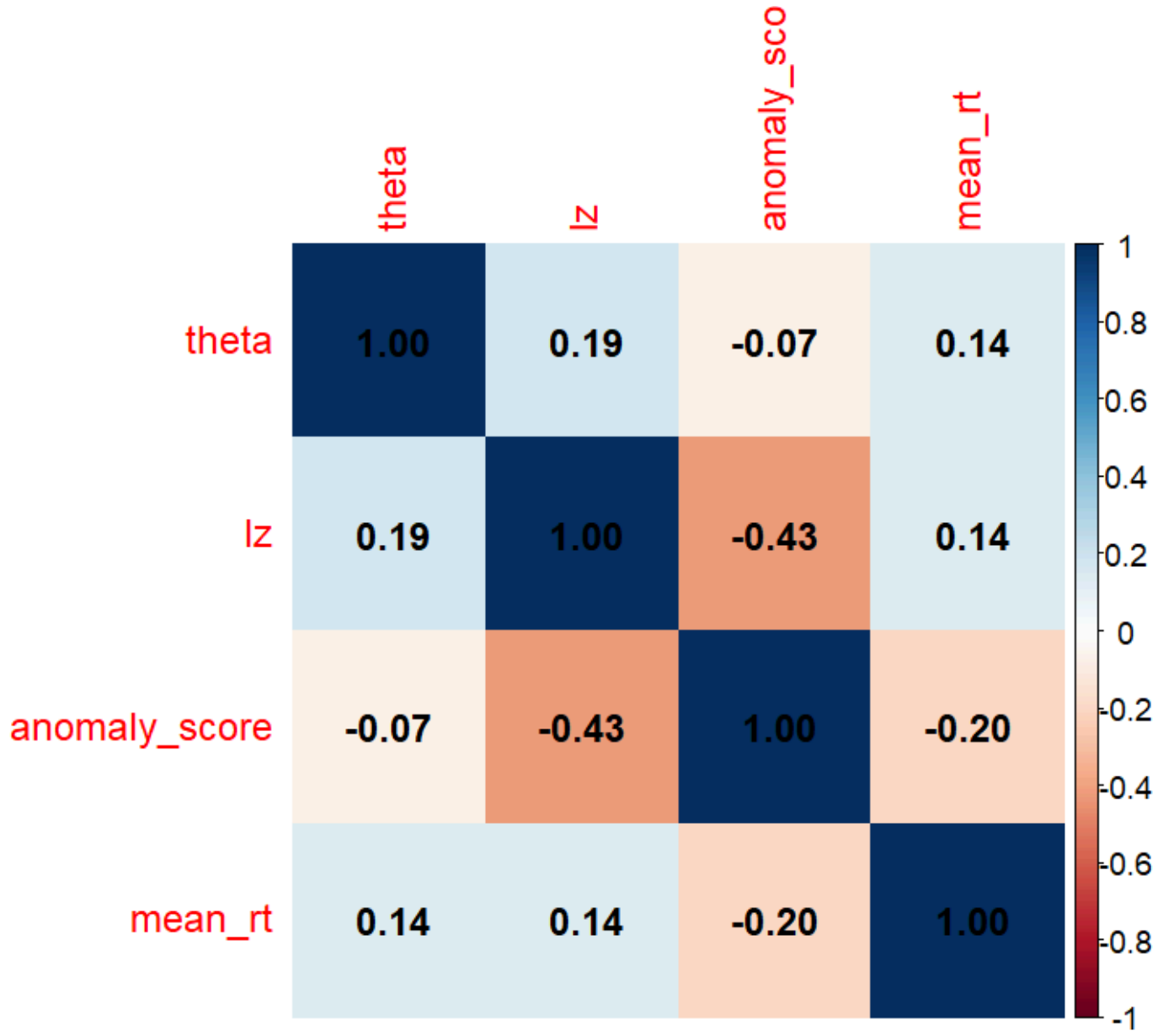


Figure 1

Correlation Matrix of Psychometric and Behavioral Features.

Correlation structure among psychometric and behavioral indicators used in the cheating-detection framework.

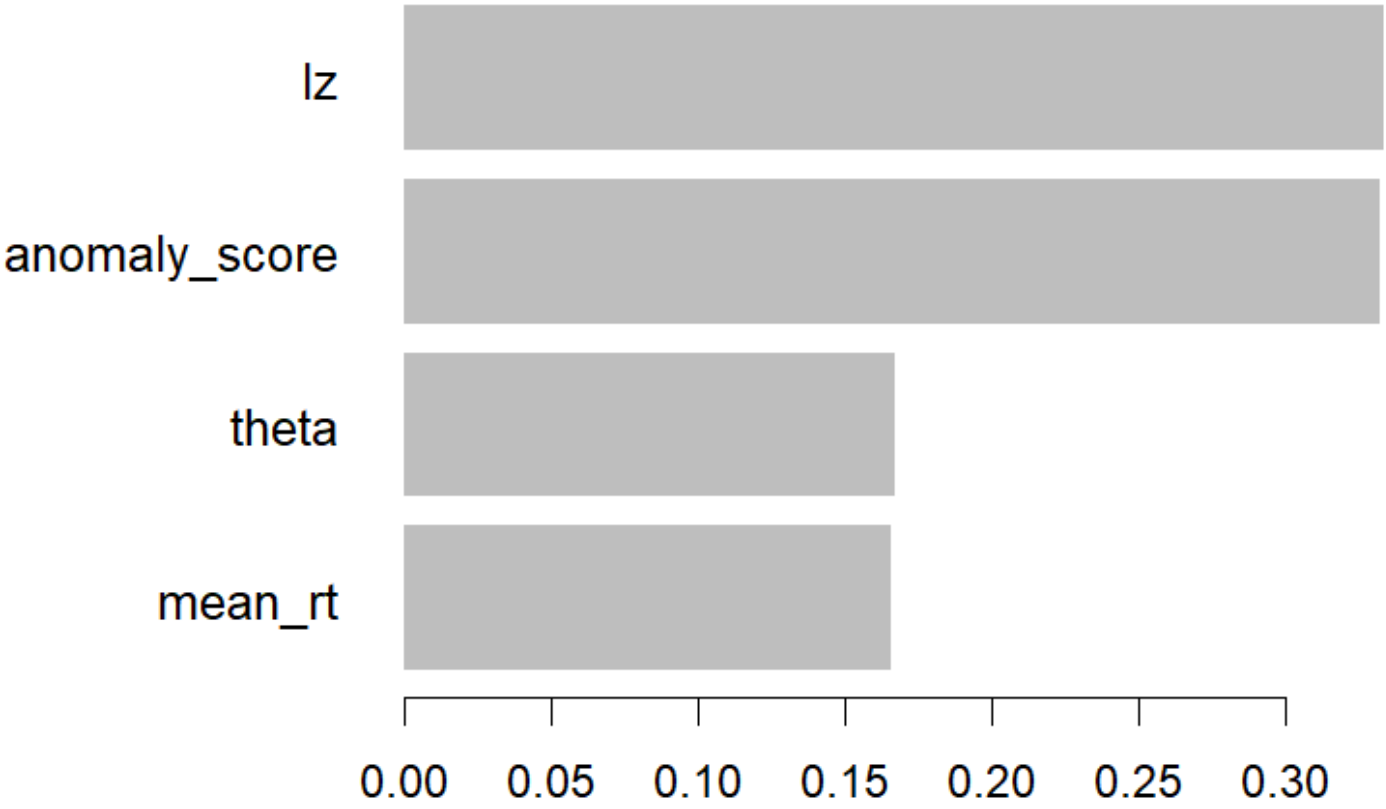


Figure 2

XGBoost Feature Importance Based on Mean SHAP Values.

Global feature contributions estimated from SHAP values in the Extreme Gradient Boosting model.

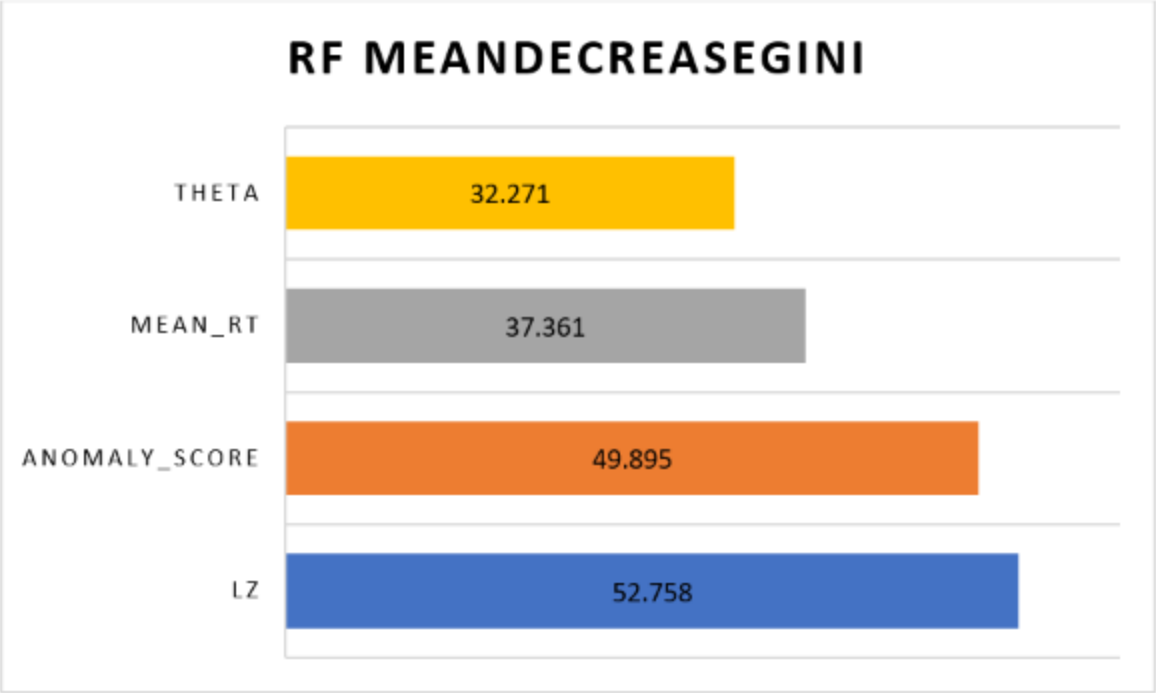


Figure 3

Random Forest Feature Importance Based on Mean Decrease Gini.
Relative importance of psychometric and behavioral predictors in the Random Forest model.

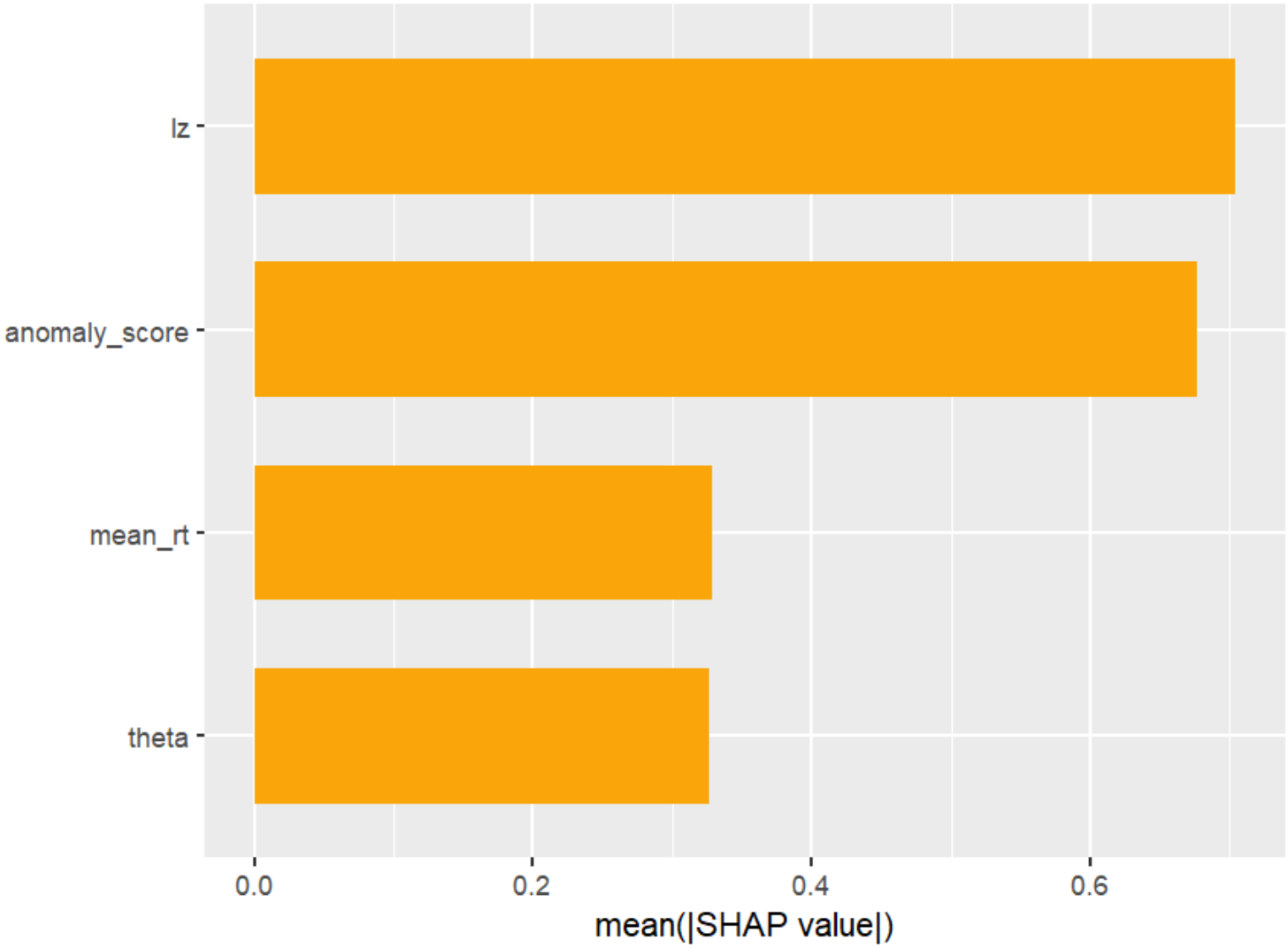


Figure 4

SHAP-Based Global Feature Importance.
Mean absolute SHAP values showing global feature contributions in the XGBoost model.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- RSyntax4analysis.pdf
- Simulationv4FinalEIT.xlsx