

A Multimodal Framework for Real-Time Detection of Computer-Based Test Malpractice in Nigeria's Unified Tertiary Matriculation Examination: A Simulation-Based Feasibility Study

Peter A. Ugwu ^{1, ✉}, Solomon K. Yateghtegh ²

1. Computer Science, Benson Idahosa University, Benin, NGA

2. Computer Science, Tansian University, Anambra, NGA

Received: August 06, 2026 | Review began: August 11, 2026 | Review ended: August 16, 2026 | Published: September 10, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Background

Nigeria's Unified Tertiary Matriculation Examination (UTME), a national computer-based test since 2015, is the admission gateway for nearly two million candidates each year. Integrity failures persist and are now technical rather than clerical. In May 2025, the Joint Admissions and Matriculation Board (JAMB) admitted that faulty server updates corrupted the responses of 379,997 candidates across 157 centres, forcing a national resit. In the same cycle, the West African Examinations Council (WAEC) withheld 192,089 results over suspected malpractice. JAMB also confirmed technology-assisted fraud, including artificial intelligence-assisted impersonation and fingerprint manipulation. Existing countermeasures, namely question randomisation, serialisation, and post-examination review, act before or after the examination rather than during it.

Objectives

This paper proposes a real-time multimodal framework that fuses three live computer-based test signals and tests the fusion principle on a synthetic benchmark.

Methods

The design specifies a vision encoder over per-workstation footage to flag gaze deviation, unauthorised devices, and multiple faces; a bidirectional long short-term memory encoder over per-item response times and keystrokes to flag pacing consistent with pre-knowledge or a relayed feed; and a graph attention network over a within-hall answer-similarity graph to flag candidate pairs whose response patterns are implausible without collusion. Since JAMB telemetry is not public, evaluation uses a synthetic benchmark of 30,000 candidate-sessions calibrated to 2024-2025 WAEC, NECO (National Examinations Council), and JAMB disclosures. This proof of concept implements the three encoders as feature groups fed to a gradient-boosted classifier and simulates the vision signal as calibrated features rather than processing real footage.

Results

On synthetic data, the feature-group proof of concept achieved an F1-score of 0.80 and an area under the receiver operating characteristic curve of 0.95, above answer-similarity (0.48), vision-only (0.42), and telemetry-only (0.33) baselines, with the largest gains on collusion and impersonation. These figures test the fusion principle on simulated signatures. They were not produced by a trained deep multimodal pipeline and are not a validated measure of real-world accuracy. Remote relay was the hardest class to detect, with an F1-score of 0.19, and the model over-flagged a synthetic accessibility sub-population at a false-positive rate of 67.5 per cent, against 1.5 per cent for the general clean population.

How to cite this article:

Ugwu P A, Yateghtegh S K (September 10, 2026) A Multimodal Framework for Real-Time Detection of Computer-Based Test Malpractice in Nigeria's Unified Tertiary Matriculation Examination: A Simulation-Based Feasibility Study. Cureus J Comput Sci 3 : es44389-026-00273-x. DOI <https://doi.org/10.7759/s44389-026-00273-x>

Conclusions

Fusing three signal groups separates simulated malpractice better than any single group. The paper sets out deployment conditions under Nigeria's power and connectivity constraints, three requirements for protecting candidates with disabilities, obligations under the Nigeria Data Protection Act 2023, and the validation using real examination data before deployment.

Categories: Ethical AI and Responsible Technology, AI applications, AI/ML-based decision support systems

Keywords: computer-based testing, examination malpractice, deep learning, anomaly detection, academic integrity

Introduction

The Unified Tertiary Matriculation Examination (UTME) is the largest standardised test in Nigeria. In 2025, 1.9 million candidates sat the examination, and their scores determined admission to universities, polytechnics, monotechnics, and colleges of education across the federation. The Joint Admissions and Matriculation Board (JAMB) made the computer-based test (CBT) compulsory for all candidates in 2015 on the argument that a fully digital, encrypted, and biometrically gated test would close the paper-based channels for leakage, impersonation, and mass copying. Ten years later, the avenues for malpractice have widened rather than narrowed.

May 2025 showed this most clearly. The JAMB registrar disclosed that an error in server updates meant that the responses of some candidates were not recorded at delivery centres over three days, affecting 379,997 candidates across 157 centres in the Lagos and south-east zones [1]. The board ordered a national resit and reopened a mop-up window for candidates who missed both sittings. It separately disclosed that some candidates had secured inflated scores by paying to join a messaging group. None of these failures was detected during the examination. Each surfaced afterwards, once scores had been released and complaints had been received.

The pattern repeats elsewhere. The West African Examinations Council (WAEC) withheld 192,089 results from candidates in the 2025 West African Senior School Certificate Examination over suspected malpractice, ranging from the use of banned mobile phones to collusion [2]. The National Examinations Council (NECO) recorded 10,094 malpractice cases in 2024 and 3,878 in 2025. JAMB has acknowledged technology-assisted fraud, including the use of artificial intelligence to impersonate candidates and the manipulation of fingerprints, though the number of confirmed cases remains small at roughly 140 out of about two million candidates, a figure that is at least as indicative of detection failure as it is of low incidence. In January 2026, the Federal Ministry of Education announced a further round of reforms for the 2026 WAEC and NECO examinations, again centred on the static pre-examination countermeasures of question randomisation and serialisation that JAMB has used since 2015 [3].

Every one of these responses acts before the candidate sits down or after the candidate leaves the hall. Each is a hardening of the server, a shuffling of the questions, or a statistical review of withheld results. None attends to what appears on the screen, in the answer log, or across the hall during the three hours of the sitting. This is the structural problem that has long troubled Nigeria's examination bodies. Malpractice is caught at a stage where the only remaining options are to withdraw a result, order a resit, or refer the matter to a disciplinary committee, all of which are disruptive, expensive, and late in protecting the candidates who did not cheat.

A parallel literature on examination integrity technology exists, but it is scattered across three tracks that rarely intersect. The first comprises answer-similarity methods, running from early copying indices through the collusion statistics of Wollack [4] to more recent machine learning approaches to test collusion. These are effective at identifying answer sheets that are suspiciously similar, but they are computed after the test has been scored and reveal nothing about candidate behaviour during the test. The second is webcam proctoring. Deep learning models trained on gaze, head pose, and object detection now achieve high accuracy in detecting visible cheating in real time, including a systematic review of 37 studies by Adeyemi et al. [5] and attention-augmented object detection for examination video reported by Zuo et al. [6]. These systems see only what the camera sees and are blind to collusion between candidates and to anomalies in pacing.

How to cite this article:

Ugwu P A, Yateghtegh S K (September 10, 2026) A Multimodal Framework for Real-Time Detection of Computer-Based Test Malpractice in Nigeria's Unified Tertiary Matriculation Examination: A Simulation-Based Feasibility Study. *Cureus J Comput Sci* 3 : es44389-026-00273-x. DOI <https://doi.org/10.7759/s44389-026-00273-x>

The third is behavioural telemetry, in particular response-time and keystroke analysis. Ranger et al. [7] showed that such analysis can identify cheaters even with no camera in the room, but telemetry alone cannot distinguish a fast and confident candidate from one typing answers fed to them from outside the room.

We found no published framework, Nigerian or international, that combines all three signal types with a real-time decision taken while the session is still running. This paper applies that fusion logic to the CBT setting and adapts it to Nigerian conditions. It makes three contributions. First, it proposes a multimodal architecture for a unified anomaly classifier that operates during the live CBT session, integrating a vision encoder, a telemetry encoder, and a graph attention collusion encoder. Second, because JAMB webcam, keystroke, and answer-log data are not public, it develops and documents a synthetic benchmark whose fraud-class prevalence is tuned to the malpractice rates reported by WAEC, NECO, and JAMB for 2024 and 2025, and it tests the fusion principle against classical and single-modality baselines on that benchmark. Third, it sets out the deployment, fairness, and legal conditions under which such a system could reasonably be piloted on real examination data in Nigeria, including the requirements of the Nigeria Data Protection Act 2023.

Materials And Methods

Fraud taxonomy

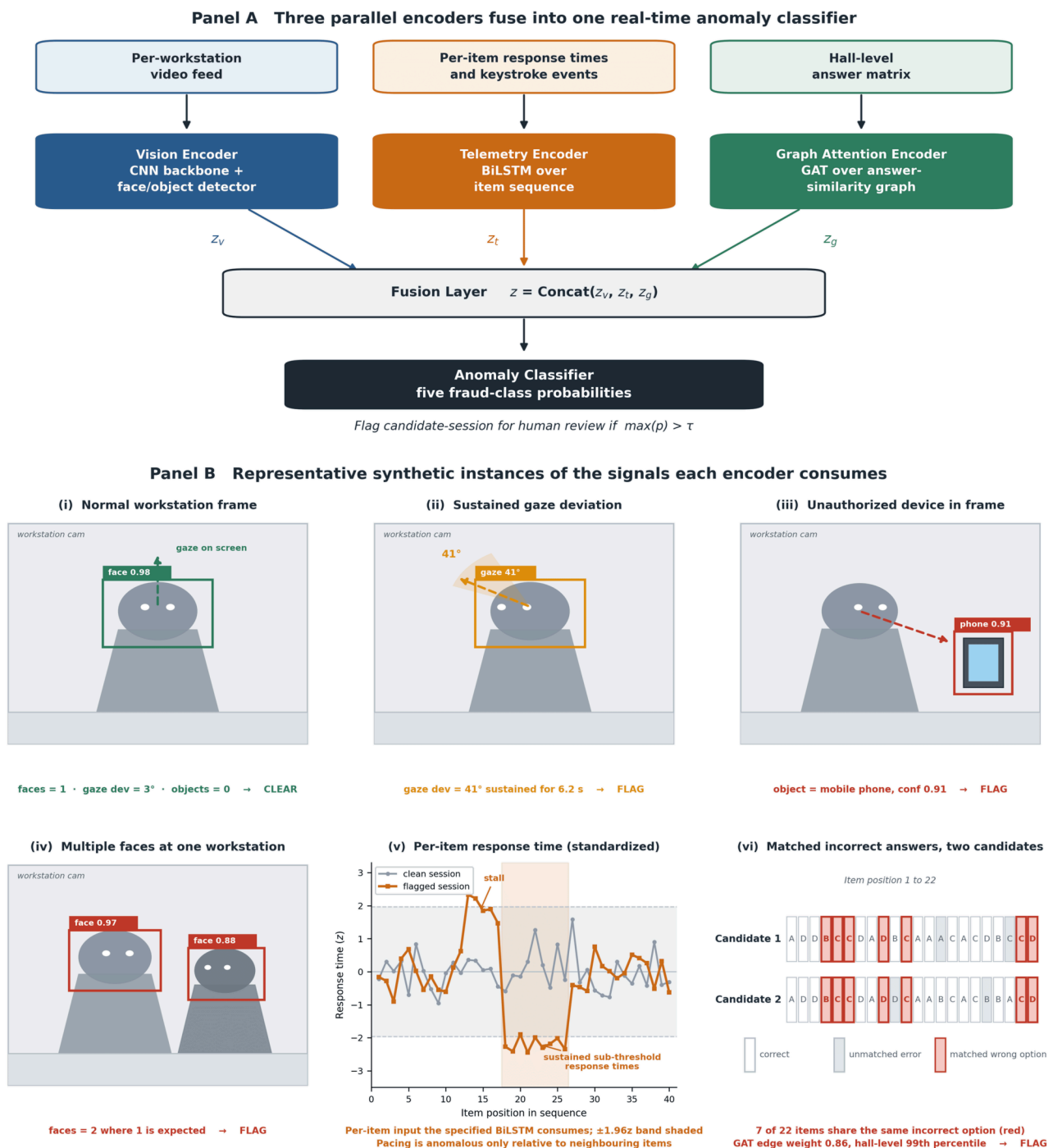
The framework covers the five classes of CBT malpractice described by WAEC, NECO, and JAMB in their public announcements for the 2024 and 2025 examinations. Impersonation is the presence at the workstation of someone other than the registered candidate. Pre-knowledge is the case of a candidate answering unusually quickly and accurately in a pattern consistent with prior sight of the questions. Collusion occurs when two or more candidates in the same session show answer patterns unlikely to arise by chance, whether through speech, hand signals, or a shared device. Unauthorised device use covers phones, smartwatches, and similar devices visible in the camera frame. Remote relay covers screen mirroring or audio relay, in which a person outside the hall supplies answers as questions are read out.

System overview

The architecture is summarised in Figure 1. Three encoders run in parallel across the data stream of each CBT session: a vision encoder over the webcam or closed-circuit feed at each workstation; a telemetry encoder over the sequence of per-item response times and keystroke or click events at that workstation; and a graph attention encoder over an answer-similarity graph built across all candidates in the same hall and session. The three embeddings are fused and passed to an anomaly classifier that returns a probability for each of the five fraud classes.

How to cite this article:

Ugwu P A, Yateghtegh S K (September 10, 2026) A Multimodal Framework for Real-Time Detection of Computer-Based Test Malpractice in Nigeria's Unified Tertiary Matriculation Examination: A Simulation-Based Feasibility Study. *Cureus J Comput Sci* 3 : es44389-026-00273-x. DOI <https://doi.org/10.7759/s44389-026-00273-x>



Panel B shows schematic instances of the signals the framework specifies at the workstation and hall level. No real footage, candidate record or identifiable person is shown. The proof-of-concept evaluation reported in this paper operates on session-level summary features derived from these signals, not on raw frames or item sequences (see Materials and Methods).

FIGURE 1: Proposed multimodal architecture for real-time CBT malpractice detection. Panel A shows the encoder and fusion architecture. Panel B shows representative schematic instances of the specified per-workstation, per-item, and hall-level signals. All Panel B instances are schematic renderings. No real footage, candidate record, or identifiable person is

How to cite this article:

Ugwu P A, Yateghtegh S K (September 10, 2026) A Multimodal Framework for Real-Time Detection of Computer-Based Test Malpractice in Nigeria's Unified Tertiary Matriculation Examination: A Simulation-Based Feasibility Study. Cureus J Comput Sci 3 : es44389-026-00273-x. DOI <https://doi.org/10.7759/s44389-026-00273-x>

shown, and the proof-of-concept evaluation operates on session-level summary features rather than raw frames or item sequences

BiLSTM, Bidirectional Long Short-Term Memory; CBT, Computer-Based Test; CNN, Convolutional Neural Network; GAT, Graph Attention Network

Vision Encoder

Video from the webcam or closed-circuit camera at each workstation is sampled and processed by a lightweight convolutional backbone combined with a face and object detector, following the approaches surveyed by Nigam et al. [8] in their review of artificial intelligence-based proctoring and implemented at scale by Lamba and Sharma [9], who fused face detection, gaze tracking, and head-pose estimation into a single behavioural score. El Kohli et al. [10] report comparable results for convolutional detection of examination fraud from classroom video. Adeyemi et al. [5] reviewed 37 studies published between 2013 and 2024 and concluded that these methods outperform human invigilation in detecting visible cheating. Zuo et al. [6] show that an attention-augmented YOLOv8 detector raises cheating-detection accuracy in examination video while remaining light enough for modest hardware. In this framework, the vision encoder produces a suspicion score per candidate per second from gaze deviation, head-pose angle, the number of faces detected in the frame, and the detection of a phone, earpiece, or paper note.

Telemetry Encoder

A second stream records, for each candidate-item pair, the response time, the number of answer changes, and the inter-keystroke or inter-click latency. In the specified design, each feature is standardised within the candidate session across the full item sequence before it enters the encoder. For candidate c , feature f , and item position i , the value is transformed to $z(c, f, i) = (x(c, f, i) - \mu(c, f)) / \sigma(c, f)$, where $\mu(c, f)$ and $\sigma(c, f)$ are the mean and standard deviation of that feature over all items the candidate attempted in that session. Where $\sigma(c, f)$ is zero, which arises when a candidate makes no answer changes at all, the feature is set to zero for that session to avoid division by zero. Within-session standardisation is specified in preference to population-level standardisation because absolute response time reflects reading speed, typing speed, and interface familiarity, all of which vary widely across candidates and none of which carries information about malpractice. Standardising within the session removes the candidate's own baseline and leaves relative deviation, which is the quantity of interest. The standardised sequence is passed to a bidirectional long short-term memory network rather than to a unidirectional network or an item-independent classifier because pacing anomalies are defined by their neighbours. A four-second response is unremarkable after a run of fast items and anomalous after a run of slow ones. A candidate receiving a relayed answer typically stalls while the request is transmitted and then produces a run of fast, correct responses, so the evidence for any single item lies in the items that follow it as well as those that precede it. A unidirectional network observes only the preceding context, and a model treating items independently observes neither. The design follows Ranger et al. [7], who compared 18 indicators of cheating in electronic examinations and found that response-time measures, such as sudden mid-test changes in pacing, discriminated between cheaters and honest test takers as well as, or better than, several older statistical measures.

The proof-of-concept evaluation reported in this paper does not execute this pipeline. It operates on four session-level summary statistics derived from the three per-item features, namely the mean and variance of response time, the number of answer changes, and the mean inter-click latency, generated on a bounded scale and passed to the gradient-boosted classifier described under Reproducibility and Experimental Environment. No item sequence, standardisation step, or trained bidirectional network is involved in the results reported below. The standardisation procedure and the bidirectional encoder set out above are specifications for a deployed implementation on real telemetry, and their evaluation is future work.

Graph Attention Collusion Encoder

Each candidate-session in a hall is represented as a node in a graph. Edge weights are computed from the number of shared incorrect or unusual answers between each pair, following the tradition of answer-similarity statistics used by Chang and Chang [11] and treated with modern machine-learning feature representations by Kamalov et al. [12], with

How to cite this article:

Ugwu P A, Yateghtegh S K (September 10, 2026) A Multimodal Framework for Real-Time Detection of Computer-Based Test Malpractice in Nigeria's Unified Tertiary Matriculation Examination: A Simulation-Based Feasibility Study. Cureus J Comput Sci 3 : es44389-026-00273-x. DOI <https://doi.org/10.7759/s44389-026-00273-x>

related gradient-boosted detection of item pre-knowledge by Zopluoglu [13]. Rather than setting a fixed threshold for each candidate pair, the framework applies a graph attention network in the manner of Veličković et al. [14], which learns to weight each candidate's neighbours and thus distinguishes a pair that shares answers from a pair that merely shares a room. Ding et al. [15] demonstrate the same mechanism for multimodal spatiotemporal anomaly detection, which is the closest published analogue to the fusion problem addressed here.

Fusion, Training Objective, and Optimisation

The three per-candidate embeddings are concatenated and passed to a five-way anomaly classifier. The design specifies training with focal loss to address the imbalance between the large number of clean candidate-sessions and the small number of injected fraud sessions, optimised using AdamW [16]. The system is a screening tool that refers cases to a human validator. It is not a disqualification system.

Three design choices deserve explicit motivation. The rationale for the bidirectional telemetry encoder is set out in the Telemetry Encoder subsection above. The graph attention network was selected over fixed pairwise thresholds because learned attention weights allow the model to separate genuine collusion from the background similarity that any two candidates in the same room share [14]. Focal loss was selected over standard cross-entropy because the injected fraud prevalence of 11.5 per cent leaves clean sessions dominating the training signal, and AdamW was selected for its decoupled weight decay [16]. Algorithm 1 gives a conceptual outline of the full screening loop for one hall-session.

Algorithm 1: Real-time fused screening for one CBT hall-session, as specified. The proof-of-concept evaluation reported below substitutes feature groups and a gradient-boosted classifier for the encoders in steps 2, 3, 5, and 7.

Input: per-workstation video frames V , per-item telemetry sequences T , and the hall answer matrix A .

Output: per-candidate fraud probabilities p over the five fraud classes.

1. For each screening interval during the live session, do the following.
2. Compute $z_v = \text{VisionEncoder}(V)$, giving gaze, head pose, face count, and object signals.
3. Compute $z_t = \text{BiLSTM}(T)$, giving pacing and keystroke dynamics.
4. Build $G = \text{BuildSimilarityGraph}(A)$, with edges weighted by shared unusual answers.
5. Compute $z_g = \text{GAT}(G)$, giving the hall-level collusion embedding.
6. Concatenate $z = \text{Concat}(z_v, z_t, z_g)$.
7. Compute $p = \text{Classifier}(z)$, trained with focal loss and optimised with AdamW.
8. If $\max(p)$ exceeds the decision threshold, flag the candidate-session for human review.
9. End for.

Synthetic Benchmark Construction

JAMB webcam video, keystroke logs, and hall-level answer matrices are not public, and no ethical clearance for their use was in place for this study. Following the approach of Zhang et al. [17] in election forensics, where synthetic data is constructed to mimic the properties of the real data it stands for, the benchmark simulates the scale of a Nigerian CBT operation at 30,000 candidate-sessions, equivalent to 120 hall-sessions of 250 candidates each. Five fraud classes were injected at rates set slightly above the 9.75 per cent malpractice rate disclosed by WAEC, giving a combined injected prevalence of 11.5 per cent so that minority classes contain enough cases for evaluation. Clean candidate-session behaviour is sampled from parameter ranges observed in the proctoring and CBT usability literature cited above, and injected fraud behaviour is drawn from distributions designed to reproduce the qualitative characteristics reported for

How to cite this article:

Ugwu P A, Yateghtegh S K (September 10, 2026) A Multimodal Framework for Real-Time Detection of Computer-Based Test Malpractice in Nigeria's Unified Tertiary Matriculation Examination: A Simulation-Based Feasibility Study. Cureus J Comput Sci 3 : es44389-026-00273-x. DOI <https://doi.org/10.7759/s44389-026-00273-x>

each class. Pre-knowledge distributions show unusually consistent per-item response times, and collusion distributions show elevated answer overlap concentrated within physically adjacent seating blocks. This is a proof-of-concept simulation. It does not use candidate data, and the results below should be interpreted accordingly.

Baselines

Three baselines are compared against the fused framework. The first is a classical answer-similarity index computed on the synthetic hall-level answer matrix alone, in the tradition of the collusion statistics of Wollack [4]. The second is a telemetry-only classifier using the telemetry feature group alone. The third is a vision-only classifier using the vision feature group alone, similar in spirit to the systems of Lamba and Sharma [9] and the LSTM-based cheating classifier of Alsabhan [18].

Evaluation metrics

Detection performance is reported as precision, recall, F1-score, and AUROC on a held-out 20 per cent split of the synthetic benchmark at the F1-optimal operating threshold. Because a CBT screen must raise a flag before the session ends, two scoring costs are also reported: amortised batch throughput and single-session latency. Both reflect the classifier's own cost and exclude the upstream video and telemetry processing that a full deployment would require.

Reproducibility and experimental environment

In this proof-of-concept release, the vision, telemetry, and graph encoders are implemented as feature groups fed into a gradient-boosted classifier rather than as trained deep networks, and the vision signal is simulated as calibrated per-candidate features rather than processed from real webcam footage. This substitution runs on a modest CPU in minutes and tests the central claim that fusing three signal groups outperforms any single group while keeping the benchmark fully synthetic. Hyperparameter settings appear in Table 7. No grid search or tuning was performed. A full deep implementation would tune the encoder and fusion hyperparameters, and that tuning is left to future work on real data. The hardware, software versions, random seeds, and evaluation protocol are given under Experimental Setup.

The benchmark generator, the models, the fixed seeds, and the full set of 30,000 candidate-sessions are publicly available under the MIT License at <https://github.com/favorproject/cbt-malpractice-detection>, release v1.0.1.

How to cite this article:

Ugwu P A, Yateghtegh S K (September 10, 2026) A Multimodal Framework for Real-Time Detection of Computer-Based Test Malpractice in Nigeria's Unified Tertiary Matriculation Examination: A Simulation-Based Feasibility Study. *Cureus J Comput Sci* 3 : es44389-026-00273-x. DOI <https://doi.org/10.7759/s44389-026-00273-x>

Setting	Value	Note
Classifier	HistGradientBoosting	CPU proof-of-concept stand-in for the deep fusion head
Max iterations	200	Fixed default; no tuning
Learning rate	0.1	scikit-learn default
Max tree depth	6	Fixed default; no tuning
Class weighting	Balanced	Handles 11.5 per cent fraud prevalence; stands in for focal loss
Test split	20 per cent	Stratified by fraud class
Random seeds	42, 43, 44, 45, 46	Five fixed values for repeated runs
Runs per result	5	One per seed; results averaged

TABLE 1: Hyperparameter settings for the proof-of-concept classifier

Results And Discussion

Experimental setup

All experiments ran on one machine: a multi-core x86-64 CPU with 8 GB of RAM running Ubuntu 24.04 LTS, with no GPU. The software stack was Python 3.12.3 with scikit-learn 1.8.0, NumPy 2.4.4, pandas 3.0.2, and SciPy 1.17.1. These version strings were captured from pip freeze at run time rather than reported from memory. Each is a released version listed on the Python Package Index, and all are pinned in the requirements file in the public repository so that a reviewer can reproduce the environment exactly.

The synthetic benchmark of 30,000 candidate-sessions was split 80 per cent to 20 per cent into training and held-out test sets, stratified by fraud class so that every fraud class appears in the test set at its injected prevalence within a combined fraud rate of 11.5 per cent. The full pipeline, from session generation through training to evaluation, was repeated five times under the fixed seeds 42, 43, 44, 45, and 46. Every figure reported below is the mean over those five runs, with the sample standard deviation given where informative. Within each seed, the fused framework and the three baselines were trained and evaluated on identical splits, so the seed-level comparisons between them are paired. Detection figures in Table 2 use the F1-optimal threshold. The ablation in Table 4 uses the fixed default threshold of 0.50, with an F1-optimal cross-check reported alongside it.

Synthetic fraud class distribution

After calibration, the number of synthetic injections per fraud class in the held-out test split is shown in Figure 2. Collusion is the largest class, consistent with the WAEC disclosure identifying collusion as its most frequently cited malpractice offence.

How to cite this article:

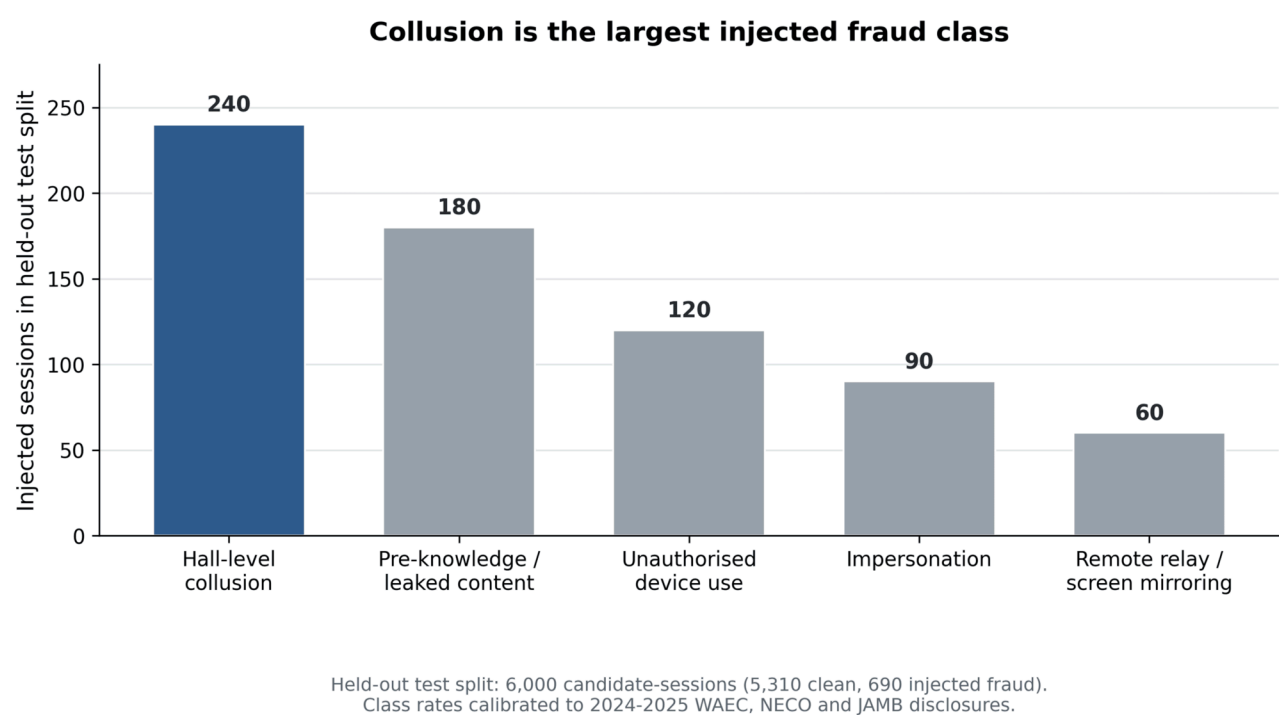


FIGURE 2: Distribution of synthetic fraud-class injections in the held-out test split

JAMB, Joint Admissions and Matriculation Board; NECO, National Examinations Council; WAEC, West African Examinations Council

Detection performance

Detection results for the fused framework and the three baselines are reported in Table 2 and Figure 3. Each value is the mean over five independent seeded runs. The fused framework achieved an F1-score of 0.80 (sample standard deviation 0.008) and an AUROC of 0.95 (sample standard deviation 0.006). The classical answer-similarity index scored 0.48 and 0.66, the vision-only baseline scored 0.42 and 0.66, and the telemetry-only baseline scored 0.33 and 0.64 on F1 and AUROC, respectively. At its F1-optimal operating point, the fused framework reached a precision of 0.86 and a recall of 0.74 on the general test population. This balance does not extend to the accessibility sub-population examined below, where the false-positive rate reaches 67.5 per cent. A paired-samples t-test on the five seed-level F1 values confirms that the fused framework outperformed the vision-only baseline, with a mean paired difference of 0.38 in F1: $t(4) = 76.70$, $p < 0.001$, Cohen's $d = 34.30$, and a 95 per cent confidence interval for the mean difference of 0.366 to 0.394.

How to cite this article:

Method	Precision	Recall	F1-score	AUROC
Answer-similarity index (Wollack-style)	0.87	0.33	0.48	0.66
Vision-only (feature stand-in)	0.77	0.29	0.42	0.66
Telemetry-only (feature stand-in)	0.39	0.29	0.33	0.64
Proposed fused framework	0.86	0.74	0.80	0.95

TABLE 2: Detection performance on the synthetic CBT benchmark, held-out test split. Values are the mean over five seeded runs

AUROC, Area Under the Receiver Operating Characteristic Curve; CBT, Computer-Based Test

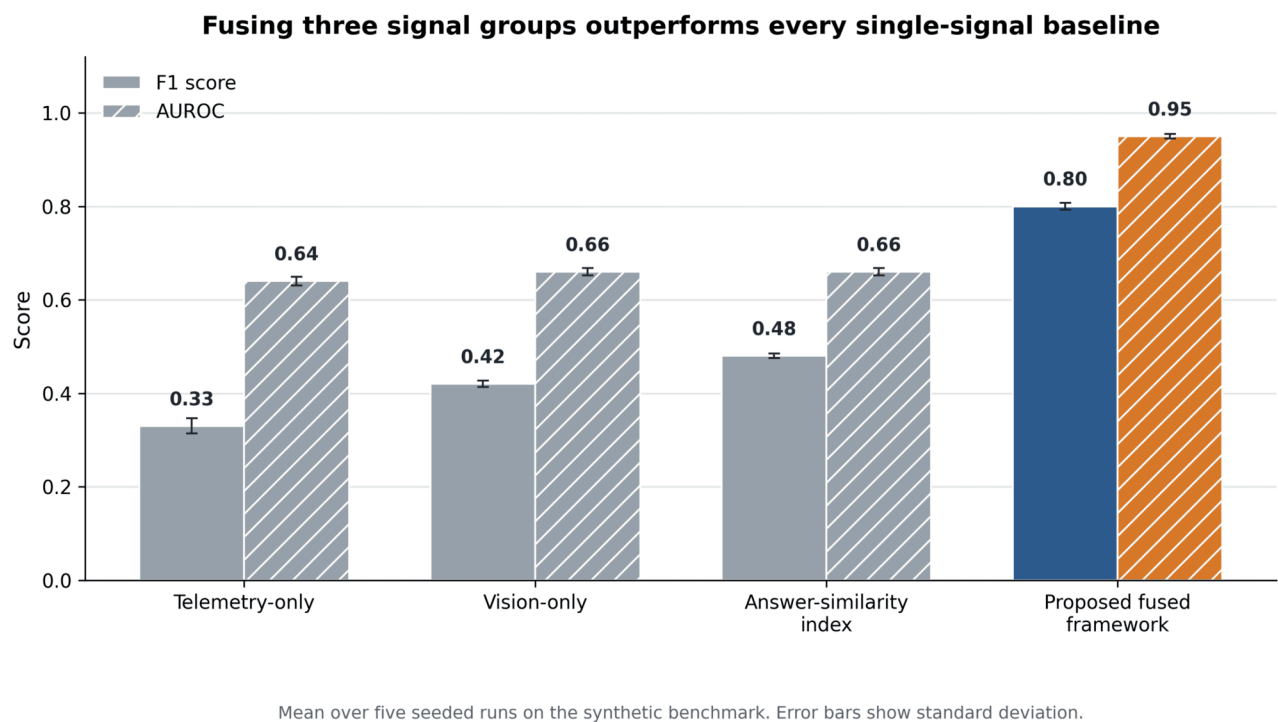


FIGURE 3: Detection performance comparison across methods on the synthetic benchmark

AUROC, Area Under the Receiver Operating Characteristic Curve

Per-class performance

Per-class scores for the fused framework are shown in Table 3. Hall-level collusion was the easiest class to detect ($F1 = 0.83$) because the graph signal captures shared-answer structure that no single workstation view can. Impersonation followed ($F1 = 0.78$), with vision and telemetry signals reinforcing each other, and unauthorised device use scored 0.69. The two hardest classes were pre-knowledge ($F1 = 0.44$) and remote relay ($F1 = 0.19$). Remote relay was the hardest of the five, which matches expectations, since a well-executed relay leaves the thinnest trace in any single signal. These two weak classes are the clearest targets for future work on richer telemetry features.

Fraud class	Precision	Recall	F1-score
Hall-level collusion	0.74	0.95	0.83
Impersonation	0.70	0.89	0.78
Unauthorised device use	0.57	0.88	0.69
Pre-knowledge or leaked content	0.30	0.82	0.44
Remote relay or screen mirroring	0.11	0.61	0.19

TABLE 3: Per-class detection performance of the proposed fused framework

Ablation study

The contribution of each modality is quantified in Table 4 and Figure 4. Every configuration in the ablation, including the full framework, is scored at the fixed default decision threshold of 0.50, which places the full framework at an $F1$ of 0.74. This is the same trained model reported in Table 2, where the $F1$ -optimal threshold raises its $F1$ to 0.80. The ablation uses 0.74 as its reference value so that all four configurations are compared at one operating point and no ablated configuration is compared against a differently thresholded baseline. Measured against that reference, removing the graph attention feature group caused the largest drop, to 0.51, a fall of 0.23. This confirms that hall-level answer structure carries a fraud signature that neither vision nor telemetry captures alone, with the biggest effect on collusion. Removing the vision group lowered $F1$ to 0.56, a fall of 0.18, reflecting the visible-behaviour information underlying device use and impersonation. Removing the telemetry group had the smallest effect, lowering $F1$ to 0.64, a fall of 0.10, with its main contribution concentrated in the pacing-driven pre-knowledge class. Every modality contributes, and the fused model beats each single-modality baseline, but the graph and vision signals carry the most weight. As a robustness check, repeating the ablation at the $F1$ -optimal threshold places the full framework at 0.80 and the losses at 0.22 for the graph group, 0.18 for the vision group, and 0.09 for the telemetry group, so the ranking and magnitude of the modality contributions do not depend on the threshold choice. Both sets of figures are reported in the released results file.

How to cite this article:

Configuration	F1-score	Change from full framework
Full framework (all three modalities), reference	0.74	-
Without graph attention group	0.51	-0.23
Without vision group	0.56	-0.18
Without telemetry group	0.64	-0.10

TABLE 4: Ablation study showing the contribution of each modality. Every configuration, including the full framework, is scored at the fixed default decision threshold of 0.50

Table 2 reports the same fused model at its F1-optimal threshold, where the F1-score increases to 0.80.

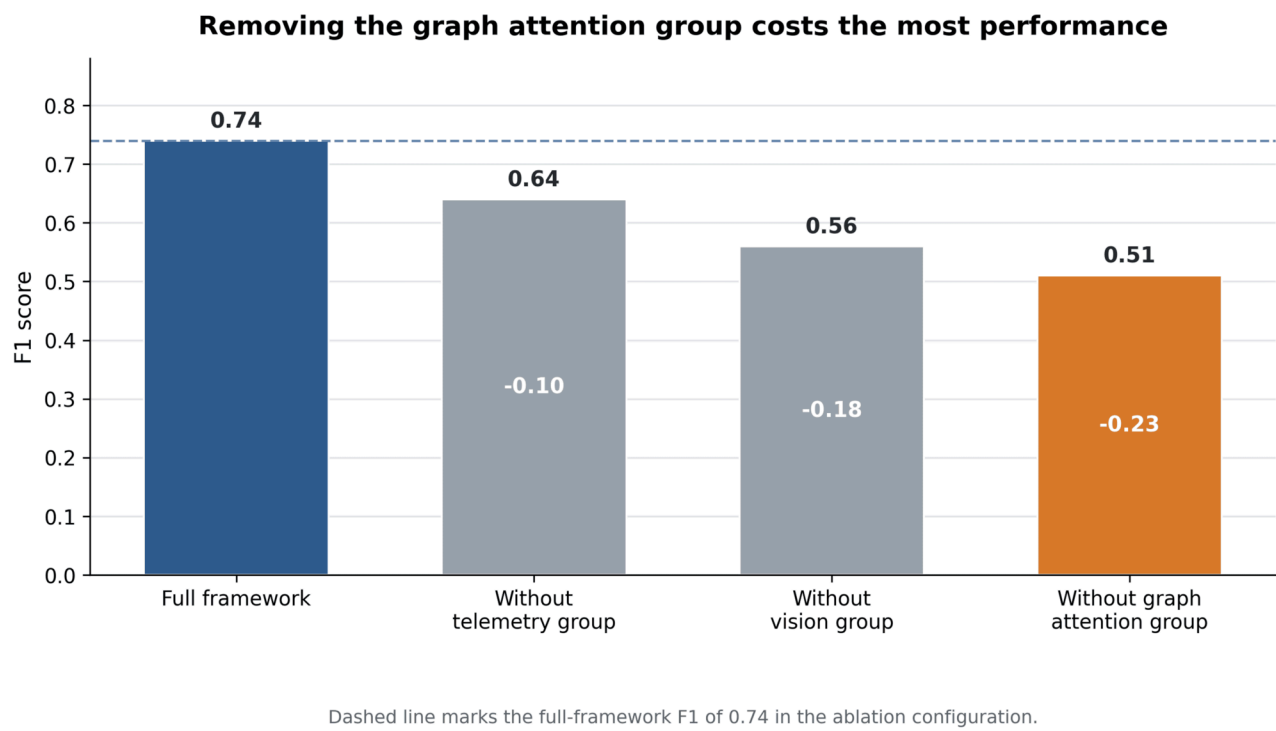


FIGURE 4: Ablation study by modality, showing the drop in F1-score when each signal is removed

Scoring latency and operational feasibility

Two scoring costs are reported because they answer different questions. Amortised batch throughput, obtained by scoring the full held-out test split in a single call on the CPU described in the Experimental Setup section and dividing by the number of sessions, averaged 0.0038 milliseconds per session. Single-session latency, obtained by scoring one candidate-session per call and averaging over 200 calls, averaged 1.2153 milliseconds per session. The first is the realistic

cost of screening a hall in bulk, and the second is the worst-case scenario in which sessions arrive one at a time. A 250-candidate hall is screened in under one second under either measure, so the classifier itself adds no meaningful delay, and a single worker can serve a full hall. Both figures cover the scoring step only. They exclude the upstream video and telemetry processing that a full deployment would require, which dominates real end-to-end latency and must be measured on actual centre hardware before any deployment claim is made.

Fairness check

The vision and telemetry signals are sensitive to atypical gaze, head movement, and pacing, so a fairness check measured how often the fused model wrongly flags clean candidates whose behaviour resembles common accessibility needs, namely slow and irregular pacing and reduced gaze stability. On this synthetic accessibility sub-population, the false-positive rate reached 67.5 per cent, compared with 1.5 per cent for the general clean population. This gap is large, and it is the central caution of this work. A model tuned only to separate fraud from typical clean behaviour treats atypical but honest behaviour as suspicious, which would turn an integrity tool into a source of harm for candidates with disabilities. The gap is a measured result rather than an incidental finding, and the discussion below sets three conditions, including a documented accommodation pathway, as hard requirements for any deployment.

Discussion

The simulation results suggest that combining visual, telemetry, and hall-level graph signals separates simulated CBT malpractice more effectively than any single signal alone. The fused framework reached an F1-score of 0.80, well above the classical answer-similarity index at 0.48, the vision-only baseline at 0.42, and the telemetry-only baseline at 0.33, and the gain over the vision-only baseline was significant across seeds. This gap matters for what an examination body can practically do with a result. An index calculated after the examination supports only a post-mortem withdrawal of results, as JAMB and WAEC did in 2025. A flag raised while a candidate is still answering can be escalated to an invigilator or the session marked for human certification during the sitting itself.

The ablation results carry a concrete operational interpretation. The vision signal does most of the work for the two classes that an invigilator could, in principle, detect by observation: device use and impersonation. JAMB's own account of the cases it detected in 2025, including device use and fingerprint manipulation, suggests that entry-point biometric checks are being actively circumvented rather than passively failing. The telemetry signal contributes most to detecting pre-knowledge, a class that leaves no visible trace and that a camera-only system, however well built, cannot structurally detect. The graph signal is the key element for collusion, the single largest failure mode of camera-only systems, and it addresses that limitation directly: a camera monitoring one workstation cannot determine whether two candidates sitting two seats apart are colluding unless it is informed about what they wrote.

Deployment feasibility must be assessed against Nigeria's existing CBT infrastructure rather than an idealised data centre. JAMB and the tertiary institutions that run their own post-UTME screening operate hundreds of centres of varying vintage, many with inadequate webcams and network connections and unreliable mains power and generator backup, compared with the better-resourced settings of most published webcam-proctoring studies. The measured scoring latency shows that the classifier is inexpensive enough to run in a zonal office without a national cloud, but that figure excludes video processing, which dominates real end-to-end cost. Any deployment claim must be verified on actual centre hardware running the full pipeline rather than inferred from classifier scoring time or from detector throughput reported on other hardware [6].

Capturing webcam images, keystroke telemetry, and hall-level answers from Nigerian candidates, most of whom are minors or newly adult, constitutes processing under the Nigeria Data Protection Act 2023. That Act requires registration of the data controller or processor with the Nigeria Data Protection Commission, a lawful basis for processing, and data minimisation. A workable pilot would need a retention period short enough to avoid holding video of the large majority of candidates who did not cheat and long enough to allow review of flagged sessions. The collusion-detection literature reviewed above also suggests that informing candidates in advance that results will be checked for collusion is among the most effective deterrents available.

How to cite this article:

Ugwu P A, Yateghtegh S K (September 10, 2026) A Multimodal Framework for Real-Time Detection of Computer-Based Test Malpractice in Nigeria's Unified Tertiary Matriculation Examination: A Simulation-Based Feasibility Study. *Cureus J Comput Sci* 3 : es44389-026-00273-x. DOI <https://doi.org/10.7759/s44389-026-00273-x>

The fairness result is not a side issue. At the F1-optimal operating threshold, the fused model flagged 67.5 per cent of the synthetic accessibility sub-population, compared with 1.5 per cent of the general clean population. The size of the gap follows from how the sub-population was constructed and from what the model learned. Candidates who need accommodations were simulated with slower and more variable pacing and reduced gaze stability. Those shifts move in the same directions as the signatures the model associates with remote relay, which raises response-time variance and inter-click latency, and with impersonation, which raises gaze deviation. A model trained only to separate injected fraud from typical clean behaviour has no basis for distinguishing atypical honest behaviour from fraud because no such examples appear in its training distribution. The result is therefore a property of the training design rather than an artefact of the simulation, and the same failure would be expected of any model trained in the same way on real data. Applied without safeguards, such a system would flag roughly two out of every three candidates in this group in error and would convert an integrity measure into a source of disadvantage for the candidates least able to contest it.

Three conditions follow, and this paper treats them as requirements rather than recommendations. A documented accommodation pathway, such as extra time or a human-proctor override, must accompany any pilot. A flag from this framework must never function as an automated determination of malpractice, and human adjudication must precede any effect on a candidate's result. Labelled accessibility profiles must be included in the training set so that the model learns to separate atypical honest behaviour from fraud rather than treating the two alike.

Four limitations restrict these results. First, the evaluation uses synthetic data built to reflect documented statistics, not actual JAMB, WAEC, or NECO records. The scores measure how well the model separates the synthetic fraud signatures on which it was calibrated. They are not a validated measure of real-world accuracy and should not be interpreted as such. Second, this proof of concept implements the three encoders as feature groups fed into a gradient-boosted classifier rather than as the trained deep networks specified by the design, and the vision signal is simulated as calibrated features rather than processed from real footage. The result confirms the design principle that three signals outperform one, but a real deployment would require the full deep pipeline trained on real data. Third, the graph collusion encoder requires an accurate seating map linking candidates, workstations, and halls. Where seating records are unreliable, the collusion signal falls back on the classical pairwise index it is intended to surpass. Fourth, the framework flags sessions for human review. It is not designed to disqualify a candidate without human oversight, given the scale of JAMB's 2025 technical failure and the harm that a wrongful accusation could cause in a national examination.

The proposed framework builds on three bodies of work. It extends the classical test-security statistics literature [4,11-13], which applies a similarity signal in isolation as a post hoc test, by embedding that signal within a real-time multimodal pipeline. It builds on the webcam-proctoring literature [5,6,9,18] by adding response telemetry and hall-level graph signals that are unavailable from a camera alone. It adds to the behavioural-telemetry literature [7], which relies on response-time evidence alone, by combining that evidence with visual and structural signals. This combination matters more than pacing by itself, since Susnjak and McIntosh [19] show that large language models are already undermining the value of any single integrity signal taken in isolation.

Implications and future directions

Three recommendations follow. JAMB, WAEC, and NECO should pilot a real-time behavioural and telemetry layer alongside existing entry-point biometric checks and post hoc statistical review, beginning with a limited number of CBT centres and a candidate-consent and data-protection framework that meets the requirements of the Nigeria Data Protection Act 2023. Tertiary institutions that run their own post-UTME screening are well placed to host smaller pilots, since they possess both the CBT infrastructure and the ethics-review structures needed to work with real, consented candidate data. Future work should extend the collusion graph from the single-hall level to the zonal level, address the central constraint posed by the absence of a real Nigerian benchmark through anonymised and ethically cleared webcam, telemetry, and answer-log data supplied by an examination body, and test the framework's robustness against the evasion strategies JAMB has already confirmed are in use, including artificial intelligence-assisted impersonation and fingerprint manipulation.

How to cite this article:

Ugwu P A, Yateghtegh S K (September 10, 2026) A Multimodal Framework for Real-Time Detection of Computer-Based Test Malpractice in Nigeria's Unified Tertiary Matriculation Examination: A Simulation-Based Feasibility Study. *Cureus J Comput Sci* 3 : es44389-026-00273-x. DOI <https://doi.org/10.7759/s44389-026-00273-x>

Conclusions

This study designed a real-time multimodal framework for detecting CBT malpractice in Nigeria's UTME and evaluated the fusion principle on a synthetic benchmark of 30,000 candidate-sessions calibrated to public WAEC, NECO, and JAMB disclosures. The fused feature groups reached an F1-score of 0.80 and an AUROC of 0.95, outperforming every single-signal baseline, with the strongest results on collusion and impersonation, the two classes most prominently highlighted by JAMB and WAEC during the 2025 examination cycle. The framework also exposed two limitations that matter more than the headline score. Remote relay was difficult to detect, with an F1-score of 0.19, and the model over-flagged a synthetic accessibility sub-population, producing a false-positive rate of 67.5 per cent. Both findings point directly to the work required before any deployment.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Peter A. Ugwu, Solomon K. Yateghtegh

Acquisition, analysis, or interpretation of data: Peter A. Ugwu, Solomon K. Yateghtegh

Drafting of the manuscript: Peter A. Ugwu

Critical review of the manuscript for important intellectual content: Peter A. Ugwu, Solomon K. Yateghtegh

Supervision: Peter A. Ugwu, Solomon K. Yateghtegh

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with [the ICMJE uniform disclosure form](#), all authors declare the following:

- **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work.
- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
- **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The data (and/or code) supporting this study are openly available in GitHub at <https://github.com/favorproject/cbt-malpractice-detection>, reference v1.0.1.

Acknowledgements

Solomon K. Yateghtegh reviewed the manuscript, provided critical feedback on the framework design, and approved the final version for submission.

How to cite this article:

Ugwu P A, Yateghtegh S K (September 10, 2026) A Multimodal Framework for Real-Time Detection of Computer-Based Test Malpractice in Nigeria's Unified Tertiary Matriculation Examination: A Simulation-Based Feasibility Study. Cureus J Comput Sci 3 : es44389-026-00273-x. DOI <https://doi.org/10.7759/s44389-026-00273-x>

References

1. Joint Admissions and Matriculation Board: Press conference address on the complaints about the 2025 Unified Tertiary Matriculation Examinations (UTME) results by the Registrar, Joint Admissions and Matriculation Board (JAMB), Prof. Is-Haq Olanrewaju Oloyede, CON, FNAL, at the boardroom of JAMB National Headquarters, Bwari, Abuja, on Wednesday, May 14, 2025. (2025). Accessed: August 11, 2026: https://jamb.gov.ng/Bulletin/2025/JAMBulletin_19-05-2025.pdf.
2. WAEC withholds 192,000 results over exam malpractice. Guardian Nigeria. (2025). Accessed: August 11, 2026: <https://guardian.ng/education/waec-withholds-192000-results-over-exam-malpractice/>.
3. Federal Ministry of Education: FG introduces new measures to eliminate WAEC, NECO examination malpractice in 2026. Federal Ministry of Information and National Orientation. (2026). Accessed: August 11, 2026: <https://fmino.gov.ng/fg-introduces-new-measures-to-eliminate-waec-neco-examination-malpractice-in-2026/>.
4. Wollack JA: A nominal response model approach for detecting answer copying. Applied Psychological Measurement. 1997, 21:307-320. [10.1177/01466216970214002](https://doi.org/10.1177/01466216970214002)
5. Adeyemi JO, Ogunlere SO, Akwaronwu BG: Real-time detection of examination malpractices using convolutional neural networks and video surveillance: a systematic review with meta-analysis. British Journal of Computer, Networking and Information Technology. 2025, 8:15-50. [10.52589/bjcnit-qc5eelje](https://doi.org/10.52589/bjcnit-qc5eelje)
6. Zuo Y, Chai SS, Goh KL: Cheating detection in examinations using improved YOLOv8 with attention mechanism. Journal of Computer Science. 2024, 20:1668-1680. [10.3844/jcscsp.2024.1668.1680](https://doi.org/10.3844/jcscsp.2024.1668.1680)
7. Ranger J, Schmidt N, Wolgast A: The detection of cheating on e-exams in higher education—the performance of several old and some new indicators. Frontiers in Psychology. 2020, 11:568825. [10.3389/fpsyg.2020.568825](https://doi.org/10.3389/fpsyg.2020.568825)
8. Nigam A, Pasricha R, Singh T, Churi P: A systematic review on AI-based proctoring systems: past, present and future. Education and Information Technologies. 2021, 26:6421-6445. [10.1007/s10639-021-10597-x](https://doi.org/10.1007/s10639-021-10597-x)
9. Lamba S, Sharma N: Deep learning-based multimodal cheating detection in online proctored exams. Journal of Electrical Systems. 2024, 20:7375-7383.
10. El Kohli S, Jannaj Y, Maanan M, Rhinane H: Deep learning: new approach for detecting scholar exams fraud. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2022, XLVI-4/W3-2021:103-107. [10.5194/isprs-archives-xlvi-4-w3-2021-103-2022](https://doi.org/10.5194/isprs-archives-xlvi-4-w3-2021-103-2022)
11. Chang SC, Chang KL: Cheating detection of test collusion: a study on machine learning techniques and feature representation. Educational Measurement: Issues and Practice. 2023, 42:62-73. [10.1111/emip.12538](https://doi.org/10.1111/emip.12538)
12. Kamalov F, Sulieman H, Santandreu Calonge D: Machine learning based approach to exam cheating detection. PLoS ONE. 2021, 16:e0254340. [10.1371/journal.pone.0254340](https://doi.org/10.1371/journal.pone.0254340)
13. Zopluoglu C: Detecting examinees with item preknowledge in large-scale testing using extreme gradient boosting (XGBoost). Educational and Psychological Measurement. 2019, 79:931-961. [10.1177/0013164419839439](https://doi.org/10.1177/0013164419839439)
14. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y: Graph attention networks. Proceedings of the 6th International Conference on Learning Representations. 2018, 1-12.
15. Ding C, Sun S, Zhao J: MST-GAT: A multimodal spatial-temporal graph attention network for time series anomaly detection. Information Fusion. 2023, 89:527-536. [10.1016/j.inffus.2022.08.011](https://doi.org/10.1016/j.inffus.2022.08.011)
16. Loshchilov I, Hutter F: Decoupled weight decay regularization. Proceedings of the 7th International Conference on Learning Representations. 2018, 1-18.
17. Zhang M, Alvarez RM, Levin I: Election forensics: using machine learning and synthetic data for possible election anomaly detection. PLoS ONE. 2019, 14:e0223950. [10.1371/journal.pone.0223950](https://doi.org/10.1371/journal.pone.0223950)
18. Alsabhan W: Student cheating detection in higher education by implementing machine learning and LSTM techniques. Sensors. 2023, 23:4149. [10.3390/s23084149](https://doi.org/10.3390/s23084149)
19. Susnjak T, McIntosh TR: ChatGPT: the end of online exam integrity?. Education Sciences. 2024, 14:656. [10.3390/educsci14060656](https://doi.org/10.3390/educsci14060656)

How to cite this article:

Ugwu P A, Yateghtegh S K (September 10, 2026) A Multimodal Framework for Real-Time Detection of Computer-Based Test Malpractice in Nigeria's Unified Tertiary Matriculation Examination: A Simulation-Based Feasibility Study. Cureus J Comput Sci 3 : es44389-026-00273-x. DOI <https://doi.org/10.7759/s44389-026-00273-x>