International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 38 العدد
Part 2 المجلد

A Proof-of-Concept Lightweight MediaPipe and ExtraTrees-Based Framework for Potential Suspicious Gaze and Head-Pose Cue Detection Using the Columbia Gaze Dataset

http://www.doi.org/10.62341/istj-vol38-2-rh64

Received	**2026/06/05**	**تم استلام الورقة العلمية في**
Accepted	**2026/06/28**	**تم قبول الورقة العلمية في**
Published	**2026/06/30**	**تم نشر الورقة العلمية في**

A Proof-of-Concept Lightweight MediaPipe and ExtraTrees-Based Framework for Potential Suspicious Gaze and Head-Pose Cue Detection Using the Columbia Gaze Dataset

Rafea Mohammed Almejrab*, Mofeda Abdelrrazik Abdelwakil Madhi, Nourah Mohammed Abdulmjid, Fatima Mohamed Hassn

College of Computer Technology - Benghazi, Libya
*Corresponding author Email: rafeaalmejrab@gmail.com
Co-authors emails: fafymady807@gmail.com, noorsoliman57@gmail.com, famaaljbyry@gmail.com

Abstract:

Automated online proctoring has become a practical requirement in remote and blended assessment, yet many systems depend on continuous connectivity, heavy processing, or intrusive data collection. This paper presents a proof-of-concept lightweight decision-support framework for detecting potential suspicious gaze and head-pose cues using MediaPipe-derived facial and iris features with an ExtraTrees decision layer. The study uses the Columbia Gaze dataset as a controlled proxy for off-camera gaze and non-frontal head-pose cues; it does not claim that such cues directly prove academic misconduct. A compact nine-feature representation is evaluated under a unified subject-independent protocol with validation-only threshold selection. On 1,260 held-out test images, ExtraTrees achieved 79.13% accuracy, 76.42% balanced accuracy, 62.50% macro F1, and 79.69% suspicious-cue recall. Because the dataset is highly imbalanced, an always-suspicious baseline achieves higher conventional accuracy but fails completely on

A Proof-of-Concept Lightweight MediaPipe and ExtraTrees-Based Framework for Potential Suspicious Gaze and Head-Pose Cue Detection Using the Columbia Gaze Dataset

http://www.doi.org/10.62341/istj-vol38-2-rh64

normal recall; therefore, balanced accuracy, macro F1, normal recall, and false-alarm rate are treated as the primary interpretation metrics. The best model still produced a 26.85% false-alarm rate, confirming that the method should be used only as a triage signal requiring temporal confirmation and human review. The study provides a reproducible low-resource baseline and identifies the main steps needed before deployment: exam-like video validation, camera calibration, threshold calibration, and human-in-the-loop review.

Keywords: online proctoring; MediaPipe; ExtraTrees; gaze cue detection; head pose; lightweight machine learning; computer vision; Columbia Gaze

إطار إثبات مفهوم خفيف قائم على MediaPipe و ExtraTrees لاكتشاف مؤشرات محتملة للنظر ووضعية الرأس المشبوهة باستخدام مجموعة بياناتColumbia Gaze

د .رافع محمد المجراب*، مفيدة عبد الرزاق عبد الوكيل ماضي، نورة محمد عبد المجيد، فاطمة محمد حسن

كلية تقنيات الحاسوب – بنغازي، ليبيا

البريد الإلكتروني للمؤلف المسؤول rafeaalmejrab@gmail.com

الملخص

أصبحت المراقبة الإلكترونية للامتحانات مطلباً عملياً في بيئات التعليم عن بعد والتعليم المدمج، إلا أن كثيراً من الأنظمة يعتمد على اتصال مستمر بالإنترنت أو معالجة ثقيلة أو جمع واسع للبيانات .تقدم هذه الورقة إطار إثبات مفهوم خفيفاً لدعم القرار في اكتشاف مؤشرات محتملة للنظر ووضعية الرأس المشبوهة، اعتماداً على ميزات الوجه والقزحية المستخرجة من MediaPipe وطبقة قرار من نوع . ExtraTrees استخدمت الدراسة

A Proof-of-Concept Lightweight MediaPipe and ExtraTrees-Based Framework for Potential Suspicious Gaze and Head-Pose Cue Detection Using the Columbia Gaze Dataset

http://www.doi.org/10.62341/istj-vol38-2-rh64

مجموعة بيانات Columbia Gaze كبيئة مضبوطة لتمثيل مؤشرات النظر بعيداً عن الكاميرا ووضعية الرأس غير الأمامية، ولا تفترض أن هذه المؤشرات وحدها تثبت سوء السلوك الأكاديمي .تم تقييم تمثيل جدولي مكون من تسع ميزات ضمن بروتوكول مستقل حسب الأشخاص، مع اختيار العتبة على مجموعة التحقق فقط .على 1,260 صورة اختبارية حقق نموذج ExtraTrees دقة 79.13% ، ودقة متوازنة76.42% ، و Macro F1مقدارها62.50% ، واستدعاء لمؤشرات الاشتباه .79.69% وبسبب عدم توازن البيانات، فإن خط أساس يتنبأ دائماً بالفئة المشبوهة يحقق دقة تقليدية أعلى لكنه يفشل كلياً في تمييز الفئة الطبيعية؛ لذلك تُعد الدقة المتوازنة و Macro F1 واستدعاء الفئة الطبيعية ومعدل الإنذار الخاطئ هي المقاييس الأهم في تفسير النتائج . تؤكد النتائج أن الإطار المقترح مناسب كأداة فرز أولية ودعم قرار، وليس كأداة اتهام تلقائي، وأنه يحتاج إلى تحقق زمني ومراجعة بشرية قبل أي تطبيق فعلي.

الكلمات المفتاحية :المراقبة الإلكترونية؛ MediaPipe؛ ExtraTrees ؛ مؤشرات النظر؛ وضعية الرأس؛ التعلم الآلي الخفيف؛ الرؤية الحاسوبية؛.Columbia Gaze

1. Introduction

Online assessment has expanded access to education, but it has also increased the need for transparent and scalable approaches to exam-integrity review. Fully manual review of recorded exams is expensive and slow, while fully automated proctoring can raise privacy, fairness, and due-process concerns when its decisions are opaque or over-sensitive. These concerns are especially important in low-resource educational settings where students may use mid-range laptops, ordinary webcams, unstable internet connections, and variable lighting conditions.

This paper focuses on a narrow visual cue: whether the test taker appears to look toward the screen/camera region or repeatedly away from it. A central clarification in this revision is that off-camera gaze is not treated as proof of cheating. A student may look away while thinking, reading from the screen below the camera, or reacting to

http://www.doi.org/10.62341/istj-vol38-2-rh64

environmental distraction. The proposed model therefore detects potential suspicious gaze/head-pose cues, not academic misconduct itself.

The proposed system is intentionally lightweight. MediaPipe is used to extract facial and iris landmarks, and a compact feature vector is classified using rule-based methods and shallow tree ensembles. This design avoids the cost and opacity of end-to-end video models and makes the decision layer easier to reproduce, inspect, and deploy on CPU-based devices.

The main contributions are:

1) a proof-of-concept MediaPipe/ExtraTrees framework for potential gaze/head-pose cue detection;
2) a subject-independent evaluation protocol on the Columbia Gaze dataset;
3) explicit treatment of class imbalance, including a majority-class baseline;
4) a detailed description of the nine features, model hyperparameters, threshold-selection rule, and evaluation metrics;
5) a human-in-the-loop deployment interpretation that prevents frame-level gaze cues from being used as standalone misconduct judgments.

2. Related Work

Automated online proctoring systems typically combine multiple cues, such as identity verification, face presence, gaze estimation, audio monitoring, screen activity, and human review. Atoum et al. [1] showed that practical exam proctoring is usually a multimedia analytics problem rather than a single-cue classification task. Recent work on AI-assisted gaze detection has similarly positioned gaze analysis as a support signal for proctors, not a replacement for human judgment [2].

http://www.doi.org/10.62341/istj-vol38-2-rh64

The Columbia Gaze dataset was introduced for gaze and eye-contact research and provides controlled combinations of head pose and gaze direction [3]. It is useful for proof-of-concept evaluation because it offers repeatable gaze/head-pose conditions, but it is not a real online-exam dataset. This distinction is important for construct validity: a controlled off-camera gaze label should be interpreted as a proxy cue, not as direct evidence of suspicious exam behavior.

MediaPipe Face Mesh and related face-landmark tools provide dense facial landmarks from monocular camera input [4], [5]. Compared with Haar Cascade face detection [6], landmark-based approaches can support feature engineering around iris position and relative face geometry. However, landmark reliability may be affected by lighting, eyeglasses, camera angle, and partial occlusion.

Tree-based ensembles remain useful when the feature space is small and tabular. Random Forests reduce variance by aggregating many decision trees [8], while Extremely Randomized Trees add stronger randomization in split selection [7]. This study uses ExtraTrees as a lightweight decision layer on MediaPipe-derived features; it does not claim superiority over deep gaze-estimation models such as MobileNetV2, EfficientNet, OpenFace, or transformer-based approaches [13], [21]-[23].

The eye-based liveness detection study by Alfagi et al. [14] used eye-based liveness detection and age estimation with ordinary camera input to build a controllable environment. The present work differs in target task: it addresses potential gaze/head-pose cues for exam-review triage rather than liveness or age estimation. Nevertheless, both studies are relevant to low-resource face and eye analysis.

3. Materials and Methods

3.1 Dataset and construct-validity framing

The experiments use the Columbia Gaze dataset, which contains 5,880 images from 56 subjects with controlled head-pose and gaze-direction combinations. The dataset was selected because it

A Proof-of-Concept Lightweight MediaPipe and ExtraTrees-Based Framework for Potential Suspicious Gaze and Head-Pose Cue Detection Using the Columbia Gaze Dataset

http://www.doi.org/10.62341/istj-vol38-2-rh64

provides structured gaze/head-pose variation and supports subject-independent splitting. The labels in this paper are therefore operational labels for potential gaze/head-pose cues, not ground-truth labels of cheating or misconduct. Table 1 reports the resulting class distribution, while Table 2 defines the exact Normal / Potential suspicious cue mapping used in the experiments.

Figure 1 summarizes the revised interpretation of the pipeline. The system converts controlled gaze/head-pose information into a frame-level cue, then uses the classifier output only as a triage flag for later review.

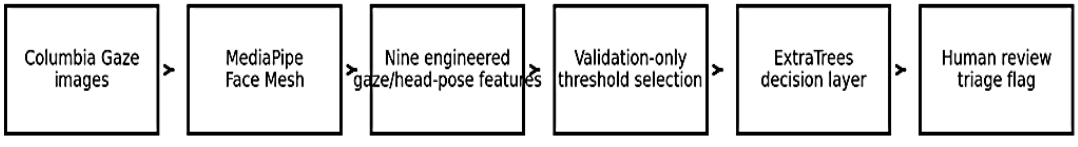

Fig. 1. Revised proof-of-concept pipeline. The model detects potential gaze/head-pose cues and does not make autonomous misconduct decisions.

TABLE 1. Dataset distribution after binary operational mapping

Class	Images	Percentage	Operational interpretation
Normal	504	8.57%	Frontal or near-screen/camera gaze cue
Potential suspicious cue	5,376	91.43%	Off-camera gaze or non-frontal pose cue
Total	5,880	100%	All images used in the experiment

TABLE 2. Explicit binary mapping rule used for Normal / Potential suspicious cue labels

Head-pose condition	Gaze-direction condition	Assigned label	Rationale
Frontal or near-frontal	Toward camera/screen region	Normal	Represents the intended on-screen attention cue

A Proof-of-Concept Lightweight MediaPipe and ExtraTrees-Based Framework for Potential Suspicious Gaze and Head-Pose Cue Detection Using the Columbia Gaze Dataset

http://www.doi.org/10.62341/istj-vol38-2-rh64

Head-pose condition	Gaze-direction condition	Assigned label	Rationale
Frontal or near-frontal	Clearly away from camera/screen region	Potential suspicious cue	Represents off-screen gaze that may require review if persistent
Non-frontal head pose	Any gaze direction	Potential suspicious cue	Large head rotation may indicate attention outside the screen area
Ambiguous or extraction failure	Unavailable/unstable landmarks	Excluded or handled as low-confidence in deployment	Not treated as positive evidence without human review

3.2 Subject-independent split

To reduce identity leakage, the final protocol used a subject-independent split. The implemented script uses GroupShuffleSplit with random_state = 42. The first split reserves 20% of subjects for testing, and the second split reserves 25% of the remaining subjects for validation. This produced 33 training subjects, 11 validation subjects, and 12 test subjects. All reported method comparisons were computed on the same held-out test split. Table 3 summarizes the subject and image counts used for training, validation, and testing.

TABLE 3. Subject-independent split distribution

Split	Subjects	Images	Normal	Potential suspicious cue	Purpose
Training	33	3,465	297	3,168	Fit the ML decision layer
Validation	11	1,155	99	1,056	Select rule and probability thresholds
Test	12	1,260	108	1,152	Final fair comparison
Total	56	5,880	504	5,376	Full dataset

A Proof-of-Concept Lightweight MediaPipe and ExtraTrees-Based Framework for Potential Suspicious Gaze and Head-Pose Cue Detection Using the Columbia Gaze Dataset

http://www.doi.org/10.62341/istj-vol38-2-rh64

3.3 Feature extraction, normalization, and engineered features

MediaPipe face and iris landmarks were converted into a compact feature vector. Horizontal head rotation was represented by yaw, while iris displacement was normalized relative to eye-region geometry so that the decision layer used relative gaze cues rather than raw pixel positions. Pitch was excluded from the final feature set because preliminary experiments showed a systematic offset that caused unstable rule-based decisions. Missing or invalid feature values were filled with zero after numeric conversion, and absolute-value features were used to make left/right and up/down deviations comparable. Table 4 lists the nine features used by the machine-learning decision layer and their expected role.

TABLE 4. Nine-feature representation used by the ML decision layer

No.	Feature	Formula / construction	Expected role
1	yaw	Estimated horizontal head rotation angle	Directional head-pose cue
2	abs_yaw	|yaw|	Magnitude of head rotation
3	iris_h	Normalized horizontal iris displacement	Left/right gaze cue
4	abs_iris_h	|iris_h|	Magnitude of horizontal gaze deviation
5	iris_v	Normalized vertical iris displacement	Up/down gaze cue
6	abs_iris_v	|iris_v|	Magnitude of vertical gaze deviation
7	iris_sum_abs	abs_iris_h + abs_iris_v	Overall iris-deviation magnitude
8	yaw_iris_h	abs_yaw x abs_iris_h	Interaction between head yaw and horizontal gaze
9	yaw_iris_v	abs_yaw x abs_iris_v	Interaction between head yaw and vertical gaze

A Proof-of-Concept Lightweight MediaPipe and ExtraTrees-Based Framework for Potential Suspicious Gaze and Head-Pose Cue Detection Using the Columbia Gaze Dataset

http://www.doi.org/10.62341/istj-vol38-2-rh64

3.4 Baselines, models, and hyperparameters

The revision adds an always-suspicious majority-class baseline because the test set is strongly imbalanced. This baseline is important for interpretation: it obtains high conventional accuracy by predicting the majority class, but it has zero ability to recognize normal frames. Therefore, the study ranks methods primarily using balanced accuracy, macro F1, normal recall, suspicious-cue recall, and false-alarm rate rather than accuracy alone. Table 5 lists the compared methods and decision mechanisms, and Table 6 reports the main hyperparameters used for the trainable models.

TABLE 5. Compared methods and decision mechanisms

Method	Input	Decision mechanism	Training
Always Suspicious	None	Predict all images as potential suspicious cue	No
Haar_existing	Image face detector output	Existing Haar-based prediction from earlier run	No
MediaPipe_rule_existing	MediaPipe features	Previously selected fixed thresholds	No
MediaPipe_rule_tuned	MediaPipe features	Validation-tuned thresholds	No supervised model
RandomForest_v2	Nine engineered features	Random Forest + validation threshold	Yes
ExtraTrees_v2	Nine engineered features	ExtraTrees + validation threshold	Yes

TABLE 6. Main hyperparameters used in the ML decision layers

Model	n_estimators	max_depth	min_samples_leaf	class_weight	random_state	n_jobs
RandomForest_v2	700	10	3	balanced_subsample	42	-1
ExtraTrees_v2	700	10	3	balanced	42	-1

3.5 Threshold selection and evaluation metrics

Probability thresholds were selected on the validation split only. For each ML model, candidate thresholds from 0.05 to 0.95 were searched in steps of 0.01. The selection objective first kept

http://www.doi.org/10.62341/istj-vol38-2-rh64

validation candidates with false-alarm rate <= 0.20 and suspicious-cue recall >= 0.60. If such candidates existed, the threshold maximizing suspicious-cue recall, then macro F1, then balanced accuracy was selected. If no candidate met the constraints, the best available macro-F1/accuracy trade-off was used. The selected thresholds were 0.56 for RandomForest_v2 and 0.47 for ExtraTrees_v2.

The main metrics are defined as follows: false-alarm rate = FP / (FP + TN), normal recall = TN / (TN + FP), suspicious-cue recall = TP / (TP + FN), and balanced accuracy = (normal recall + suspicious-cue recall) / 2. Weighted F1 is reported for completeness but is not treated as a primary ranking metric because it is dominated by the majority class.

3.6 Statistical analysis

The reported confidence intervals are descriptive 95% bootstrap intervals generated from the held-out test predictions using 500 image-level resamples with random_state = 42. Because the Columbia Gaze dataset contains multiple images per subject, image-level bootstrap intervals may be optimistic if interpreted as subject-level inference. For this reason, the manuscript avoids strong claims of statistical superiority and treats confidence intervals as descriptive uncertainty indicators. A stricter follow-up validation should use clustered bootstrap or repeated group cross-validation at the subject level and paired comparisons between models.

4. Results

Table 7 shows the final fair comparison on the same subject-independent test split. The always-suspicious baseline demonstrates the danger of interpreting conventional accuracy under severe imbalance: it reaches 91.43% accuracy and about 87.33% weighted F1, but its normal recall is 0.00% and its false-alarm rate is 100%. In contrast, ExtraTrees improves balanced accuracy, macro F1, and normal-frame recognition while retaining useful suspicious-cue recall.

A Proof-of-Concept Lightweight MediaPipe and ExtraTrees-Based Framework for Potential Suspicious Gaze and Head-Pose Cue Detection Using the Columbia Gaze Dataset

http://www.doi.org/10.62341/istj-vol38-2-rh64

TABLE 7. Unified test-set comparison including the majority-class baseline

Method	Accuracy	Balanced Acc.	Macro F1	Weighted F1	Normal Recall	Susp. Recall	False Alarm
Always Suspicious	91.43%	50.00%	47.76%	87.33%	0.00%	100.00%	100.00%
Haar_existing	16.75%	51.11%	16.74%	17.34%	92.59%	9.64%	7.41%
MediaPipe rule existing	69.68%	72.09%	55.22%	76.31%	75.00%	69.18%	25.00%
MediaPipe rule tuned	71.67%	69.82%	55.66%	77.73%	67.59%	72.05%	32.41%
RandomForest_v2	77.46%	75.51%	61.04%	82.00%	73.15%	77.86%	26.85%
ExtraTrees_v2	79.13%	76.42%	62.50%	83.19%	73.15%	79.69%	26.85%

Figure 2 visualizes the main class-aware metrics. ExtraTrees provides the strongest balance among the tested trainable approaches, while the always-suspicious baseline confirms why accuracy and weighted F1 cannot be used alone.

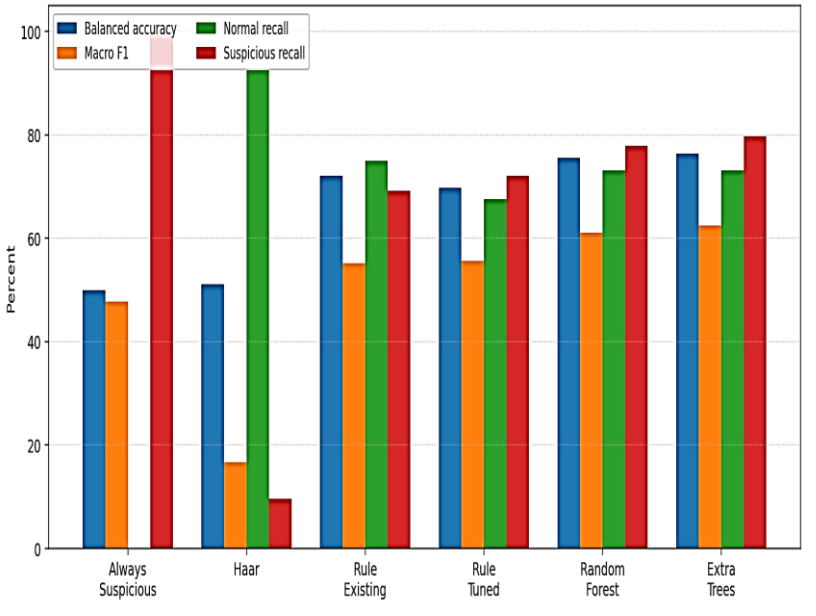

Fig. 2. Comparison of class-aware metrics. The majority-class baseline has high suspicious recall but zero normal recall.

A Proof-of-Concept Lightweight MediaPipe and ExtraTrees-Based Framework for Potential Suspicious Gaze and Head-Pose Cue Detection Using the Columbia Gaze Dataset

http://www.doi.org/10.62341/istj-vol38-2-rh64

Table 8 provides the detailed ExtraTrees confusion matrix on the held-out test split, and Figure 3 visualizes the same error pattern for easier interpretation.

TABLE 8. ExtraTrees_v2 confusion matrix on the held-out test split

Actual / Predicted	Predicted Normal	Predicted Potential suspicious cue
Actual Normal	79	29
Actual Potential suspicious cue	234	918

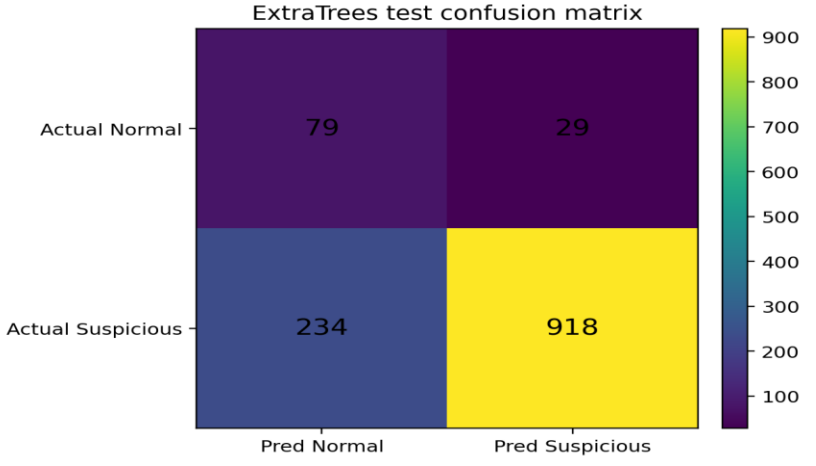

Fig. 3. ExtraTrees_v2 confusion matrix. The false-alarm rate remains too high for autonomous enforcement.

The ExtraTrees model detected 918 of 1,152 suspicious-cue test images and correctly recognized 79 of 108 normal images. Its normal recall of 73.15% is a substantial improvement over the majority baseline, but 29 normal images were still falsely flagged. This confirms the practical need for temporal smoothing and human review before any alert is acted upon.

http://www.doi.org/10.62341/istj-vol38-2-rh64

5. Discussion

The results support the limited claim that a lightweight MediaPipe/ExtraTrees pipeline can provide a reproducible frame-level cue for exam-review triage. The method is not a complete online-proctoring system. It does not verify identity, detect multiple people, monitor audio, detect phones or books, inspect browser activity, or model temporal behavior. Those tasks require a broader multimodal architecture and governance process.

The most important methodological issue is constructing validity. Columbia Gaze provides controlled gaze/head-pose conditions, but the real online-exam context is more complex. A student may look below the camera to read a laptop screen, look aside while thinking, or change posture naturally. Therefore, the paper deliberately uses the phrase potential suspicious cue and recommends that alerts be aggregated over time and reviewed by a human before any decision.

The second issue is class imbalance. In a dataset where more than 91% of images are mapped to the suspicious-cue class, accuracy can be misleading. The majority baseline exposes this problem directly. The model should therefore be judged by whether it improves class-aware metrics and reduces false alarms relative to the operational need, not by accuracy alone.

Table 9 summarizes deployment safeguards that should be applied before the current prototype is considered for real exam settings.

TABLE 9. Practical deployment recommendations

Issue	Risk	Recommended mitigation
Brief natural gaze shift	Student is flagged while thinking or reading	Use temporal smoothing over 5-10 consecutive frames before raising an event
Screen below webcam	Normal screen reading may look like downward gaze	Perform camera/screen calibration before the exam
Uneven lighting or low-quality webcam	Unstable landmarks and false alarms	Add pre-exam camera-quality and illumination checks

http://www.doi.org/10.62341/istj-vol38-2-rh64

Issue	Risk	Recommended mitigation
Eyeglasses or reflections	Eye landmarks become unreliable	Use landmark confidence and route low-confidence cases to human review
Multiple faces or missing face	Integrity cues not captured by gaze alone	Add a separate face-presence and face-count module
High false-alarm cost	Student anxiety and unfair intervention	Use human-in-the-loop review and multi-level risk labels

6. Limitations

- The evaluation uses controlled static images rather than real online-exam videos; temporal behavior is not modeled.
- The Normal / Potential suspicious cue mapping is an operational proxy and does not prove academic misconduct.
- The dataset is highly imbalanced; therefore, accuracy and weighted F1 are secondary metrics.
- The false-alarm rate of 26.85% is too high for autonomous enforcement.
- The reported bootstrap intervals are descriptive and should be replaced by clustered subject-level inference in future validation.
- Pitch was excluded because preliminary experiments showed a systematic offset.
- The study does not include direct numerical comparison with deep lightweight models such as MobileNetV2, EfficientNet, OpenFace, or compact YOLO-based modules.
- Real deployment requires exam-like videos, camera calibration, consent procedures, data-minimization rules, and appeal mechanisms.

7. Conclusion

This paper presented a proof-of-concept lightweight MediaPipe and ExtraTrees framework for detecting potential suspicious gaze and head-pose cues using the Columbia Gaze dataset. The revised

http://www.doi.org/10.62341/istj-vol38-2-rh64

analysis clarifies the operational nature of the labels, adds a majority-class baseline, documents the subject-independent protocol, lists the nine engineered features, and reports model hyperparameters and threshold selection. ExtraTrees achieved the best balance among the tested trainable methods on balanced accuracy and macro F1, but the majority-class baseline shows why conventional accuracy is not sufficient under strong class imbalance. The current framework should therefore be considered a transparent low-resource research baseline and triage module, not a standalone proctoring product. Future work should validate the model on real exam videos, use clustered subject-level statistical testing, compare against lightweight deep baselines, and integrate temporal confirmation with human review.

Data and Code Availability

The experiments use the publicly available Columbia Gaze dataset. The evaluation code records the binary label mapping, GroupShuffleSplit protocol, threshold searches, model settings, confusion matrices, and summary metrics. Any public code release should exclude private student data and include a split file to support reproducibility.

Ethical Considerations

The proposed framework should not be used to automatically accuse students of misconduct. Off-screen gaze may occur naturally during thinking, reading, stress, or environmental distraction. Appropriate use requires transparency, student notification, minimal data collection, local processing where possible, human review, and an appeal or correction mechanism.

References

[1] Y. Atoum, L. Chen, A. X. Liu, S. D. H. Hsu, and X. Liu, "Automated online exam proctoring," IEEE Transactions on Multimedia, vol. 19, no. 7, pp. 1609-1624, July 2017, doi: 10.1109/TMM.2017.2656064.

http://www.doi.org/10.62341/istj-vol38-2-rh64

[2] Y.-S. Shih, Z. Zhao, C. Niu, B. Iberg, J. Sharpnack, and M. B. Baig, "AI-assisted gaze detection for proctoring online exams," arXiv preprint arXiv:2409.16923, 2024, doi: 10.48550/arXiv.2409.16923.

[3] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking: Passive eye contact detection for human-object interaction," Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST), pp. 271-280, 2013, doi: 10.1145/2501988.2501994.

[4] Google AI Edge, "Face landmark detection guide," available online: https://developers.google.com/edge/mediapipe/solutions/vision/face_landmarker, accessed Jun. 28, 2026.

[5] Google AI Edge, "Face detector task guide," available online: https://developers.google.com/edge/mediapipe/solutions/vision/face_detector, accessed Jun. 28, 2026.

[6] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," CVPR, 2001.

[7] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Machine Learning, vol. 63, pp. 3-42, 2006.

[8] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32, 2001.

[9] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[10] G. Bradski, "The OpenCV library," Dr. Dobb's Journal of Software Tools, 2000.

[11] Z. Zhang, "A flexible new technique for camera calibration," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 11, pp. 1330-1334, 2000.

[12] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," CVPR, 2014.

http://www.doi.org/10.62341/istj-vol38-2-rh64

[13] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," IEEE FG, 2018.

[14] A. Alfagi, Z. Khlifa, M. Alrayes, and N. Omran, "Real-Time Liveness Detection Algorithm Based on Eyes Detection and Utilize Age Estimation Technique to Build a Controllable Environment," International Science and Technology Journal, vol. 30, July 2022.

[15] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," ICCV, 2007.

[16] S. Tirunagari et al., "Detection of face spoofing using visual dynamics," IEEE Transactions on Information Forensics and Security, vol. 10, no. 4, pp. 762-777, 2015.

[17] W. Kim, S. Suh, and J.-J. Han, "Face liveness detection from a single image via diffusion speed model," IEEE Transactions on Image Processing, 2015.

[18] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009.

[19] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," NeurIPS, 2012.

[20] M. Sandler et al., "MobileNetV2: Inverted residuals and linear bottlenecks," CVPR, 2018.

[21] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," ICML, 2019.

[22] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," ICLR, 2021.

[23] N. Mehrabi et al., "A survey on bias and fairness in machine learning," ACM Computing Surveys, 2021.

http://www.doi.org/10.62341/istj-vol38-2-rh64

[24] B. Efron and R. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall/CRC, 1993.

[25] D. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," Journal of Machine Learning Technologies, 2011.

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[27] Honorlock, "Exam proctoring: secure exams start to finish," available online: https://honorlock.com/why-honorlock/, accessed Jun. 28, 2026.

[28] Meazure Learning, "ProctorU to discontinue exam integrity services that rely exclusively on AI," available online: https://www.meazurelearning.com/resources/proctoru-to-discontinue-exam-integrity-services-that-rely-exclusively-on-ai, accessed Jun. 28, 2026.

[29] Moodle Plugins Directory, "Proctoring for Moodle," available online: https://moodle.org/plugins/quizaccess_proctoring, accessed Jun. 28, 2026.