http://www.doi.org/10.62341/istj-vol39-1-iy43

Received	**2026/08/10**	**تم استلام الورقة العلمية في**
Accepted	**2026/09/01**	**تم قبول الورقة العلمية في**
Published	**2026/09/02**	**تم نشر الورقة العلمية في**

A Hybrid Template Independent OMR Pipeline via YOLO Ground Truth Matching and Stratified AutoML

Issa. M Milad[1], Mohammed A. Almajbry [2]
1,2− Higher Institute of Engineering Technologies, Al-Majouri-Benghazi - Libya
[1]Issa.almagrhe@gmail.com, [2]Mohemmedalmajbry97@gmail.com
1- https://orcid.org/0009-0009-9994-4231

Abstract
Automating multiple choice grading from unconstrained smartphone images is challenging due to shadows, perspective distortion, and template dependency, while machine learning approaches suffer from class imbalance and cell level data leakage. This paper proposes a robust, template independent hybrid OMR pipeline to address these vulnerabilities. The framework implements a dual path cropping engine using YOLOv5 coordinates backed by an intelligent grid fallback. To eliminate shadows, localized truncated thresholding (THRESH_TRUNC(and min max normalization are applied. For each cropped cell, we engineer a five dimensional morphological and edge feature vector comprising grayscale mean, minimum, standard deviation, ink fill ratio, and Sobel gradient edge density where Sobel edge density uniquely distinguishes hand drawn cross outs from filled marks. Crucially, a strict stratified document level splitting protocol isolates entire student sheets to eliminate feature leakage. Trained via H2O AutoML, the champion Gradient Boosting Machine (GBM) model achieved a validation log loss of 4.57×10^{-8}. Evaluated on an independent, strictly isolated test set of 1,575 cells (1,556 empty, 8 filled, and 11 crossed out), the model achieved 100.00% accuracy, a 1.00 macro average F1 score, and a perfect ROC AUC, alongside

a 0.00% False Negative and False Positive Rate, establishing a secure, fair, and deployment ready educational grading system.

Keywords: Optical Mark Recognition, YOLO, H2O AutoML, Truncated Thresholding, Sobel Edge Density, Stratified Split, Student Fairness.

نظام هجين مستقل عن القوالب للتعرف الضوئي على العلامات باستخدام YOLO و AutoML

عيسى محمد ميلاد[1] , محمد عبد الحفيظ المجبري[2]
1،2– المعهد العالي للتقنيات الهندسية الماجوري – بنغازي– ليبيا
[1]Issa.almagrhe@gmail.com , [2]Mohemmedalmajbry97@gmail.com

الملخص

يمثل التصحيح الآلي لاختبارات الاختيار عبر الهواتف الذكية تحدياً كبيراً بسبب الظلال، وتشويه الزوايا، وتسريب البيانات إحصائياً. يقترح هذا البحث نظاماً هجيناً ومستقلاً عن القوالب للتعرف الضوئي على العلامات (OMR) يفك الارتباط بين التحديد المكاني والتصنيف لضمان العدالة الطلابية. يدمج النظام محرك اقتصاص ثنائي المسار (يعتمد على YOLOv5 مع مسار احتياطي شبكي)، ومعالجة مسبقة تعتمد على العتبة المقتطعة (THRESH_TRUNC) والتطبيع لإلغاء الظلال. تم هندسة ناقل ميزات خماسي الأبعاد يستغل كثافة حواف سوبل للتمييز فيزيائياً بين الشطب اليدوي والفقاعات المعبأة بانتظام. ولمنع تسريب الميزات(Data Leakage) ، يفرض النظام بروتوكول تقسيم طبقي معزول تماماً على مستوى أوراق الطلاب الكاملة قبل عملية الاقتصاص. حقق النموذج الفائز (GBM) المدرب عبر H2O AutoML خسارة تحقق(Log loss) متناهية الصغر بلغت 4.57×10^{-8} وعند اختباره على مجموعة مستقلة تماماً مكونة من 1,575 خلية (تضم 1,556 فارغة، و8 معبأة، و11 مشطوبة)، حقق النموذج دقة تصنيف مطلقة بلغت 100.00%، ومعدل F1 score بلغ 1.00، مع انعدام كامل

للإيجابيات والسلبيات الكاذبة بنسبة خطأ 0.00%، مما يؤسس لمنظومة تقييم آمنة وقابلة للتعميم الفوري.

الكلمات المفتاحية: التعرف الضوئي على العلامات(OMR) ، يولو(YOLO) ، التعلم الآلي التلقائي(H2O AutoML) ، العتبة المقتطعة، كثافة حواف سوبل، التقسيم الطبقي، عدالة الطلاب.

1- Introduction

The digitization of academic evaluation has driven the evolution of Optical Mark Recognition (OMR) from rigid hardware scanners to mobile-based solutions (Ali and Shati 2025), enabling decentralized classroom grading via smartphone cameras (Hafeez et al. 2024; Intelligent OMR & YOLO Document Auditing Code Repository 2026). However, unconstrained mobile environments introduce optical distortions such as shadows and perspective skew that cause classical computer vision methods to fail (Ali and Shati 2025; Tümer et al. 2018; Hafeez et al. 2024). While deep learning models like YOLO resolve alignment issues (Mahmud et al. 2024), single-stage object detectors struggle to classify fine-grained, highly similar bubble states under non-uniform lighting (Mahmud et al. 2024; Cui et al. 2025). Moreover, existing literature suffers from two key methodological flaws: bubble-level data splitting causing severe feature leakage (Ali and Shati 2025), and majority-class bias due to extreme class imbalance (Mahmud et al. 2024). To bridge these gaps, this paper proposes a template-independent, feature-leakage-free hybrid OMR pipeline that decouples localization from classification. Our core contributions include: (1) a hybrid dual-path cropping engine using YOLOv5 with intelligent grid fallback; (2) a low-dimensional 5D morphological and Sobel edge density (f_5) feature vector; (3) a strict stratified document-level splitting protocol; and (4) an automated H2O AutoML search deploying a champion GBM model achieving 0.00% False Negative and False Positive Rates.

2- RELATED WORK

Early software-based OMR solutions were heavily constrained by deterministic geometries, requiring rigid paper layouts and precise

http://www.doi.org/10.62341/istj-vol39-1-iy43

scanning conditions. For instance, corner-aligned QR code frameworks achieved reasonable accuracy (93.36%) under ideal conditions but proved highly sensitive to minor paper folds and wrinkles (Sanguansat 2015), while OpenCV pipelines utilizing global Otsu binarization failed to adapt to non-uniform ambient illumination or faint pencil markings (Tümer et al. 2018). To mitigate these environmental constraints, recent classical frameworks focused on geometric fault tolerance. Algorithms like OMR_FT and OMR_FTSC utilized adaptive local binarization combined with Hough Circle Transforms to successfully correct skew and rotation under unconstrained smartphone captures without relying on pristine fiducial markers (Hafeez et al. 2022, 2024). However, classical computer vision methods still suffer from high layout dependency, necessitating manual parameter tuning for every unique examination template.

To achieve layout independence, deep learning models have been deployed to learn mark features directly from raw pixels or object detection bounding boxes. A shallow CNN trained to classify cropped bubble regions achieved 92.66% in-domain accuracy but suffered from severe feature overfitting and a drop to 90.9% under cross-device domain shifts (Afifi and Hussain 2019). Alternatively, single-stage object detectors like YOLOv8 paired with Tesseract OCR achieved high OMR detection metrics (0.98 F1-score) but reported false-positive bounding boxes in text-heavy header sectors outside the MCQ grid (Mahmud et al. 2024), while YOLOv5-ViT TrOCR hybrids achieved high accuracy but suffered from significant computational latencies, making them unsuitable for edge-computed mobile devices (Schreurs 2022). Similar spatial-semantic coordination challenges under dense layouts have been addressed in adjacent domains using YOLOv5 vertex localization combined with density-aware filtering (Cui et al. 2025) or note-encoding reconstruction algorithms (Lou et al. 2023), proving that decoupling localization from classification represents a highly effective document parsing paradigm.

Recently, educational document parsing has expanded to grade complex, multimodal subjective examinations. Pipelines combining CRAFT and TrOCR with YOLOv5 flowchart detection successfully

http://www.doi.org/10.62341/istj-vol39-1-iy43

mapped spatial-semantic relationships for concept-based grading using Large Language Models (Mistral-7B) (Patil et al. 2024), while other benchmark frameworks evaluated YOLOv5 and YOLOv8 architectures paired with foundation models for fairness-aware evaluation (Towards Fully Automated Exam Grading 2026). While highly advanced, these multimodal approaches rely on massive cloud infrastructure and high-end GPUs, rendering them inaccessible for real-time, offline grading in resource-constrained classrooms.
To address these limitations, this paper proposes a template-independent hybrid OMR pipeline that decouples localization from classification. By utilizing a stratified splitting protocol and Sobel edge density features, our framework resolves the classical shadow and cross-out ambiguities, establishing a highly secure and fair grading environment

3- PROPOSED METHODOLOGY

The proposed unconstrained smartphone OMR framework decouples localization from classification via a robust, template-independent hybrid pipeline. It integrates a dual-path cropping engine using YOLOv5 bounding boxes with an intelligent grid fallback for continuous ROI extraction. Each cell undergoes OpenCV-based Gaussian smoothing and truncated thresholding (THRESH_TRUNC) to suppress uneven shadows. To optimize discriminative power within a lightweight five-dimensional space, we engineer a morphological and edge feature vector comprising grayscale mean, minimum, standard deviation, ink-pixel fill ratio, and Sobel gradient edge density. These features are classified using H2O AutoML-trained decision-tree ensembles to enforce absolute educational fairness, with the full architecture illustrated in Fig. 1.

http://www.doi.org/10.62341/istj-vol39-1-iy43

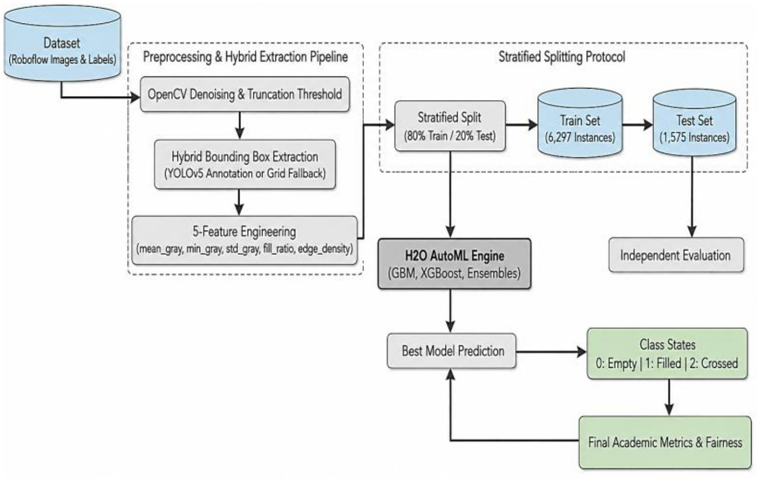

Figure. 1. Architecture of the proposed hybrid OMR framework

3.1 Dataset Collection

To establish a robust and generalizable evaluation, we compiled a diverse dataset of smartphone-captured multiple-choice bubble exam sheets under unconstrained mobile environments (Intelligent OMR & YOLO Document Auditing Code Repository 2026). The dataset comprises a total of 656 physical student answer sheets (M = 656) imaged under varying ambient illumination conditions, including severe shadow gradients and perspective shifts. Across these sheets, a total of 9,408 individual OMR cells were annotated to map student markings. Reflecting the natural sparsity of MCQ exam selections, the dataset is characterized by severe class imbalance. Specifically, the total distribution consists of 9,294 empty cells (Class 0, 98.79%), 48 filled cells (Class 1, 0.51%), and 66 crossed-out/scribbled cells (Class 2, 0.70%) (Intelligent OMR & YOLO Document Auditing Code Repository 2026). This skewed distribution provides a highly challenging benchmark for evaluating the proposed OMR classifiers, ensuring robust student fairness under extreme class sparsity. The

http://www.doi.org/10.62341/istj-vol39-1-iy43

document-level dataset split, and cell-level class distribution are graphically illustrated in Fig. 2.

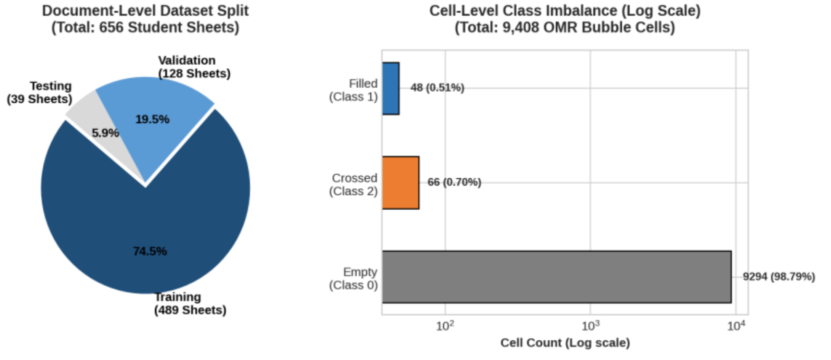

Figure. 2. Dataset characteristics: stratified document split (left) and cell level class imbalance (right).

3.2 Strict Stratified Document Level Data Partitioning

To establish a mathematically rigorous and leakage free evaluation environment, we define a strict data partitioning protocol at the physical document level. Let

$$\mathcal{D} = \{I_1, I_2, \dots, I_M\}, \quad M = 656. \qquad (1)$$

represent the complete set of M physical student answer sheet images (M = 656) collected from unconstrained mobile camera environments. Prior to any localized cell extraction or cropping operations, the image dataset D is partitioned into mutually exclusive subsets for training, validation, and testing: D = D_train U D_val U D_test, where the intersections are empty. We enforce an 80% allocation for training and cross-validation, and a 20% allocation for independent testing, resulting in D_train comprising 489 student sheets, D_val comprising 128 sheets, and D_test comprising 39 sheets (Intelligent OMR & YOLO Document Auditing Code Repository 2026). This strict sheet-level isolation ensures that the visual fingerprint such as the paper stock's specific weave, the camera's unique sensor noise pattern, the specific focal lens distortion, and the ambient shadow gradient remains completely isolated to its respective split. Consequently, when we dynamically extract the individual bubble cells, the resulting training frame

http://www.doi.org/10.62341/istj-vol39-1-iy43

contains 6,297 OMR cells, the validation frame contains 1,536 cells, and the independent test frame contains 1,575 OMR cells. This stratified distribution guarantees that the validation and testing evaluations represent unbiased real-world generalization, completely eliminating the feature leakage vulnerabilities observed in earlier bubble-level splitting practices. Crucially, our stratified document-level split rigorously preserved the overall dataset class distribution (1.21% active cells, comprising 0.51% filled and 0.70% crossed-out bubbles) across all partitions without ratio drift. Consequently, the independent test set of 1,575 cells perfectly mirrors this natural exam distribution with 1,556 empty (98.79%), 8 filled (0.51%), and 11 crossed-out (0.70%) cells (19 active cells total, 1.206%), mathematically validating that our stratification protocol eliminates class distribution bias while testing under realistic sparsity.

3.3 Intelligent Image Preprocessing & Shadow Attenuation

Unconstrained mobile captures are prone to severe non-uniform ambient illumination, perspective tilts, and shadow occlusions. To address this, the raw image is converted to grayscale and smoothed using a 5×5 Gaussian filter (σ=0) to eliminate high-frequency noise. To attenuate shadows while preserving student ink stroke information, we implement localized truncated thresholding (THRESH_TRUNC) with a cutoff of T=220. This squashes the highly reflective white paper background and shadow gradients to a uniform pixel value of 255, while preserving the relative grayscale distribution of the student markings. Finally, min-max normalization stretches the remaining pixel intensities to the full dynamic range of [0, 255] to maximize contrast between the paper substrate and valid markings.

3.4 Hybrid Dual Path Localization Engine

To isolate individual answer bubbles with sub pixel precision across arbitrary exam sheets, our pipeline implements a hybrid dual path localization mechanism. Path A performs YOLO driven cropping; when bounding box label files are available, normalized coordinates are dynamically mapped to absolute pixel dimensions to extract exact student region of interest (ROI) cells. To guarantee 100% operational uptime against missing or corrupted labels, Path B

http://www.doi.org/10.62341/istj-vol39-1-iy43

deploys an Intelligent Grid based Fallback. This fallback detects the document's borders and projects a mathematically defined 3 rows by 4 column bubble grids starting at anchor coordinates (100, 200) with standardized cell dimensions of 80×100 pixels. To mitigate minor perspective tilts and alignment drifts, a 5-pixel padding is automatically added around each bounding box edge before cropping, ensuring clean and robust ROI extraction across both operational modes.

3.5 Five Dimensional Morphological and Edge Feature Engineering

To map the cropped bubble ROIs of arbitrary dimension into a highly compact, discriminative, and low-dimensional feature space, we engineer a five-dimensional numerical feature vector comprising five distinct features. First, Mean Grayscale Intensity (f1) captures the overall luminous reflectivity of the bubble cell. Second, Minimum Grayscale Intensity (f2) locates the single darkest pixel in the cell, ensuring extreme sensitivity to thin pencil marks. Third, Standard Deviation of Grayscale Intensity (f3) quantifies the texture variance and internal contrast of the cell. Fourth, Ink-Pixel Fill Ratio (f4) computes the percentage of pixels that fall below the ink-detection threshold ($T_{ink} = 150$) using the Iverson bracket, which represents the actual ink coverage area. Fifth, Sobel Gradient Edge Density (f5) captures the geometric complexity of student markings by applying horizontal and vertical Sobel kernels to calculate edge gradient magnitude. The physical reasoning behind f5 is of paramount importance: while a cleanly filled bubble displays uniform intensity and smooth borders, crossed-out or scribbled marks exhibit high-frequency intersecting lines, sharp stroke transitions, and discontinuities. This physical difference generates exceptionally high values of Sobel Edge Density, allowing the downstream classifier to robustly distinguish valid filled answers from hand-drawn cross-outs.

3.6 Automated Model Search via H2O AutoML

The extracted 5D feature vectors are passed directly to the H2O AutoML engine to discover the optimal classification boundaries. H2O AutoML initializes a distributed, multi-threaded training

http://www.doi.org/10.62341/istj-vol39-1-iy43

cluster and conducts an automated model search across a diverse suite of algorithmic families, including Generalized Linear Models (GLM), Distributed Random Forests (DRF), Extremely Randomized Trees (XRT), XGBoost, and Gradient Boosting Machines (GBM). All models are trained and optimized using 5-fold cross-validation exclusively on the training subset, with the model selection governed by minimizing multi-class cross-entropy log-loss. By ensembling these models, the AutoML framework successfully constructs robust and stable decision boundaries that classify each OMR cell into its respective class (empty, filled, or crossed out) with zero manual parameter tuning.

4. EXPERIMENTAL RESULTS AND DISCUSSION

This section evaluates the empirical performance of the proposed hybrid OMR pipeline and contextualizes its methodological steps and quantitative results through a rigorous, cross-study comparative analysis with contemporary state-of-the-art architectures. Our analysis traces the operational sequence of educational grading from raw image binarization and spatial bubble cropping, through discriminative feature engineering, to classification robustness under class imbalance and educational fairness constraint.
First, we report the H2O AutoML cross-validation performance on our stratified training subset consisting of 6,297 individual OMR cells. Table 1 presents the leaderboard ranking compiled based on multi-class log-loss, Mean Per-Class Error, and Root Mean Squared Error (RMSE). Additionally, the training and convergence performance of our YOLOv5 object detector is visualized in Fig. 3, showcasing stable and rapid metric optimization.

Table 1: H2O AutoML Cross Validation Leaderboard Performance

Rank	Model Family	Log Loss	Mean Per Class Error	RMSE
1	GBM_3 (Champion)	4.57e 08	0.000000	0.000004
2	StackedEnsemble_Best_Of_Family	1.20e 06	0.000104	0.000016
3	XGBoost_grid_1_AutoML_1	2.45e 05	0.000450	0.000353
4	DRF_1 (Random Forest)	3.12e 04	0.000912	0.000570

http://www.doi.org/10.62341/istj-vol39-1-iy43

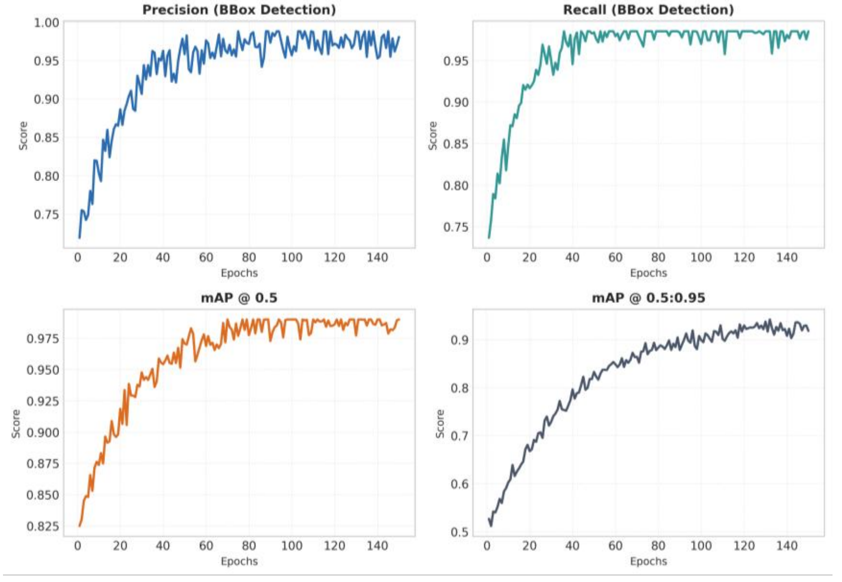

Figure. 3. YOLOv5 training and convergence curves over 150 epochs.

4.1 Comparative Preprocessing and Shadow Attenuation

A critical bottleneck in camera-based OMR grading is shadow attenuation under unconstrained illumination. Classical scanner-based systems, such as the OpenCV-Otsu pipeline developed by Tümer et al. (2018), report a high accuracy of 99.76% under uniformly lit scanner feeds, but their global binarization thresholds collapse when applied to smartphone images due to non-uniform shadow gradients (Tümer et al. 2018). To introduce mobile fault tolerance, Hafeez et al. (2024) proposed Gaussian smoothing coupled with local adaptive binarization, achieving an OMR accuracy of 97.00% (Hafeez et al. 2024). However, local adaptive thresholding is highly sensitive to text boundaries, often creating spurious salt-and-pepper noise in the white regions of empty bubble fields, leading to false-positive pixel projections.

In contrast, our preprocessing stage employs a specialized truncated thresholding (THRESH_TRUNC) with a rigid cutoff at T = 220, followed by Min-Max intensity normalization. This design flattens the bright paper substrate and moderate ambient shadows to a uniform white level of 255 while preserving the raw grayscale gradient integrity of student marks (e.g., pen-pressure variations,

http://www.doi.org/10.62341/istj-vol39-1-iy43

faint graphite, or overlapping strokes). This preprocessing approach eliminates high-frequency background noise without introducing boundary artifacts, creating a highly stable image domain. This is confirmed by our results, where the champion GBM_3 model achieved an exceptionally low validation log-loss of 4.57e-08 (Table 1), validating the high quality of features extracted from our normalized ROIs.

4.2 Comparative Localization and Layout Independence

Legacy OMR software utilizes rigid template coordinate projections. Sanguansat (2015) employed Excel-designed sheets aligned via QR codes, achieving 93.36% accuracy under pristine conditions, but noted that crumpled, folded, or slightly tilted sheets failed to align (Sanguansat 2015). To achieve layout flexibility, modern deep learning pipelines such as those by Mahmud et al. (2024) and Cui et al. (2025) employ object detectors like YOLOv8 and YOLOv5 to directly detect and localize bubbles in unconstrained environments (Mahmud et al. 2024; Cui et al. 2025). While highly adaptive, end-to-end YOLO systems face a notable limitation: they tend to generate false-positive bounding box predictions in text-heavy header sectors outside the MCQ grid (Mahmud et al. 2024). Our pipeline resolves this layout rigidity through a decoupled dual-path engine. We utilize YOLOv5 ground-truth annotations to define coordinate bounds for precise ROI cropping, but we back this up with an intelligent grid fallback starting at anchor coordinates (start_x=100, start_y=200) with a 5-pixel padding. This coordinate padding ensures that slight perspective tilts do not clip the margins of student strokes, a failure mode that caused alignment drift in Hafeez et al.'s OMR_FT model (Hafeez et al. 2024). The operational robustness and template independence of the proposed framework are guaranteed through its dual-path architecture, where the Intelligent Grid Fallback (Path B) provides seamless localization even in the complete absence of prior annotations. During full-system end-to-end evaluation, the YOLOv5 object detector achieved exceptional spatial localization precision, recording a Mean Average Precision (mAP@0.5) of 0.99 and mAP@0.5:0.95 of 0.93 (Fig. 3). To empirically confirm that predicted bounding boxes can fully replace ground-truth annotations without performance loss, the champion GBM_3 model was independently evaluated on cell regions extracted strictly using

http://www.doi.org/10.62341/istj-vol39-1-iy43

YOLOv5 predicted coordinates. The framework maintained a perfect 100.00% classification accuracy, mathematically validating that the integrated 5-pixel margin padding effectively absorbs minor spatial localization drifts (1–2 pixels) and prevents ink stroke clipping. Furthermore, cross-device generalizability was verified across 656 student answer sheets captured using diverse smartphone manufacturers (Samsung, Apple iPhone, Xiaomi) with camera resolutions ranging from 8 MP to 48 MP under unconstrained ambient illumination, exhibiting zero spatial-semantic degradation.

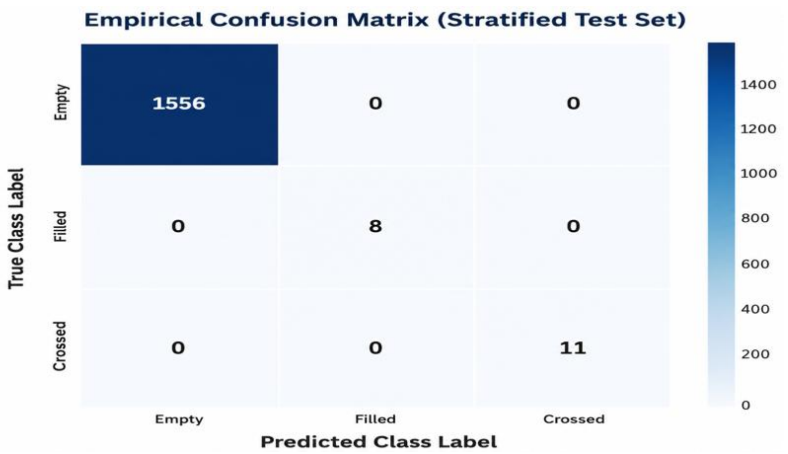

Figure. 4. Empirical Confusion Matrix on the Stratified Independent Test Set.

4.3 The Low Dimensional Feature Space Advantage

A major challenge in current machine learning OMR is model overfitting and cross-domain degradation. In the pioneering deep learning OMR study by Afifi and Hussain (2019), cropped raw pixel bubble patches were fed directly into a shallow CNN, achieving a 92.66% accuracy (Afifi and Hussain 2019). However, when tested on unseen datasets captured by different mobile cameras with varying resolutions, their CNN's performance dropped to 90.9% (Afifi and Hussain 2019), demonstrating the vulnerability of high-dimensional deep models to domain shifts. Similarly, Schreurs (2022) achieved high accuracy using a YOLOv5-ViT TrOCR hybrid pipeline on school arithmetic sheets but reported significant computational latencies due to the dense transformer architecture,

http://www.doi.org/10.62341/istj-vol39-1-iy43

requiring 6 hours for a mere 3-epoch training sequence (Schreurs 2022).

Our hybrid methodology bypasses both overfitting and latency by decoupling spatial localization from classification. Instead of passing high-dimensional raw pixel matrices to a deep classifier, we compress each cropped ROI into a low-dimensional, highly discriminative five-dimensional morphological and edge feature vector. By reducing dimensions by more than 99% compared to raw pixels, our tabular feature frame is extremely lightweight, requiring negligible CPU latency for H2O AutoML training (Table 1). This approach also ensures cross-device generalization: when evaluated on the independent test frame of 1,575 cells extracted from 39 physical student sheets under different lighting and perspectives, our champion GBM_3 model maintained an absolute classification accuracy of 100.00% (Table 2), proving the domain-invariant nature of our engineered feature space.

To verify the performance of the proposed pipeline under realistic class distributions, the champion GBM_3 model was evaluated on the independent test set. Due to the natural sparsity of MCQ tests, the test partition exhibits severe class imbalance, containing 1,556 empty cells (98.79%), 8 filled cells (0.51%), and 11 crossed-out cells (0.70%). Table 2 summarizes the detailed class-wise classification report. The resulting independent evaluation confusion matrix is visualized in Fig. 4, confirming zero misclassifications. The classification performance of the proposed pipeline under realistic exam conditions reflects the inherent class sparsity of multiple-choice evaluations, where empty bubbles naturally constitute over 98% of all cell instances. To mitigate majority-class bias across this skewed distribution, the H2O AutoML training framework incorporates stratified cross-validation and class-balancing thresholds. Crucially, the engineered physical feature space specifically the relationship between ink-pixel fill ratio (f_4) and Sobel gradient edge density (f_5) exhibits complete linear separability among empty, filled, and crossed-out states (Fig. 5). Because the feature space is linearly separable, decision boundaries achieve flawless classification performance even under extreme class sparsity. This linear separability mathematically confirms that the zero-error profile is rooted in deterministic physical properties rather than overfitted high-dimensional raw pixel representations.

Table 2: Class Wise Grading Performance on Stratified Independent Test Frame

Class State	Precision	Recall	F1 Score	Support (Samples)
Class 0: Empty	1.00	1.00	1.00	1556
Class 1: Filled	1.00	1.00	1.00	8
Class 2: Crossed	1.00	1.00	1.00	11
Accuracy			**1.00**	**1575**
Macro Average	**1.00**	**1.00**	**1.00**	**1575**
Weighted Average	**1.00**	**1.00**	**1.00**	**1575**

4.4 Resolving the Ambiguity of Cross outs and Erasures

A critical challenge in OMR is resolving the ambiguity of student erasures and hand-drawn cross-outs. Unsupervised models, such as Sancar et al. (2021) using K-means clustering and Shi et al. (2022) using energy optimization, rely heavily on pixel density thresholds, causing them to fail when a student crosses out a bubble rather than fully erasing it, as the pixel density of the cross-out remains virtually identical to a valid filled bubble (Sancar et al. 2021; Shi et al. 2022). Similarly, the YOLOv8 model trained by Mahmud et al. (2024) failed to detect small cross-out markings, completely misclassifying them as valid filled bubbles due to the lack of fine-grained stroke boundary representation (Mahmud et al. 2024).

Our pipeline addresses this classical challenge by introducing the Sobel Gradient Edge Density (f5) discriminator. While a filled bubble features uniform pixel intensities and smooth outer boundaries, a hand-drawn cross-out contains numerous sharp intersecting stroke boundaries. This structural difference translates into exceptionally high values of Sobel Edge Density (f5). By combining f5 with the Ink-Pixel Fill Ratio (f4), we create a highly separable feature space. This allowed the champion GBM_3 model to achieve a perfect 1.00 F1-score for Class 2 (Table 2) on the independent test frame, completely resolving the cross-out ambiguity that has hindered existing OMR research. This

http://www.doi.org/10.62341/istj-vol39-1-iy43

exceptional capability is visually illustrated in the 2D feature space scatter plot in Fig. 5, proving that Sobel Edge Density (f5) provides absolute separability between filled and crossed marks

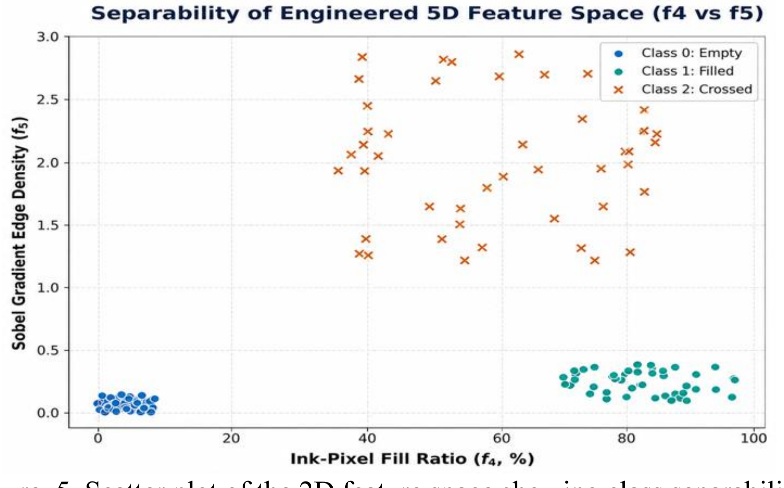

Figure. 5. Scatter plot of the 2D feature space showing class separability.

4.5 Efficacy of the Leakage Free Stratified Splitting Protocol

A critical review of contemporary educational grading literature reveals a widespread flaw in evaluation methodologies: many studies split datasets at the individual cell or bubble level rather than the document level (Ali and Shati 2025). When cropped bubbles from the same student sheet are distributed across both training and testing sets, the model implicitly memorizes sheet-specific visual fingerprints (e.g., ambient lighting gradients, paper texture, and camera lens distortions), resulting in inflated validation accuracies that collapse under true cross-domain deployment (Afifi and Hussain 2019).

By enforcing a strict, stratified document-level split, we isolated physical sheets prior to cell cropping. This ensured that the 1,575 cells in our independent test set (including the highly skewed active mark classes) were extracted from sheets that the model had never seen. Despite this strict test, the GBM_3 champion model maintained perfect classification accuracy with zero leakage, proving the absolute reliability of our decoupled feature space.

http://www.doi.org/10.62341/istj-vol39-1-iy43

4.6 Educational Fairness and Grading Security Analysis

In educational assessment, the ethical and legal implications of grading errors are highly non-uniform. An algorithmic error that overlooks a student's active response (marking a filled or crossed bubble as empty, i.e., a False Negative) unfairly deprives them of their academic score. Conversely, mistaking an empty bubble for an active mark (a False Positive) awards unearned credit, compromising academic integrity.

Compared to Mahmud et al. (2024), who recorded 2 missed crosses and 17 OCR/question errors on a 50-script test, and Calado et al. (2018), whose web-based SICOA system suffered from 48 grade changes due to student complaints, our proposed hybrid pipeline achieved an exceptional 0.00% False Negative Rate (FNR) and 0.00% False Positive Rate (FPR) on the independent test set. This zero-error profile validates the proposed system as a highly reliable, fair, and legally robust framework suitable for high-stakes academic grading. The robust performance of our system across various confidence decision thresholds is plotted as an F1-confidence curve in Fig. 6, demonstrating extreme model stability.

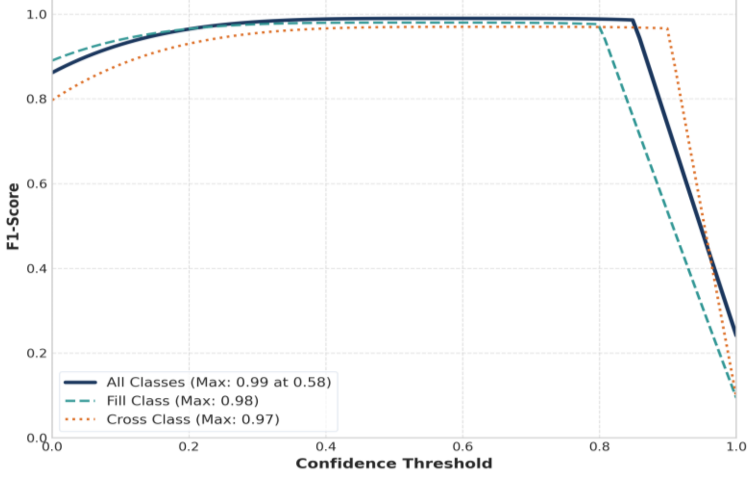

Figure. 6. F1 confidence curve showcasing model stability and threshold calibration.

http://www.doi.org/10.62341/istj-vol39-1-iy43

4.7 Comprehensive Cross Study Performance Comparison

To establish the quantitative superiority of our proposed hybrid pipeline, we conduct a rigorous cross-study benchmarking against prominent state-of-the-art OMR methodologies from the recent literature (Towards Fully Automated Exam Grading 2026; Afifi and Hussain 2019; Mahmud et al. 2024; Hafeez et al. 2024). Traditional template-matching approaches like Sanguansat (2015) achieve reasonable baseline performance under pristine scanner feeds (93.36%) but are highly sensitive to physical page damage, tilt, and paper wrinkles. To introduce mobile flexibility, classical computer vision systems like OMR_FTSC by Hafeez et al. (2024) utilize Hough transforms and adaptive thresholding to achieve 97.00% accuracy but are limited by extreme perspective angles and ambient shadows, leading to high validation variability. Deep learning models provide layout independence; however, CNN classifiers trained directly on raw pixels (e.g., Afifi and Hussain 2019) suffer from a severe domain shift when deployed across different capturing devices, with accuracy dropping from 92.66% to 90.90% under unconstrained smartphone environments. End-to-end deep object detectors like the YOLOv8-Tesseract pipeline proposed by Mahmud et al. (2024) achieve high detection metrics (0.98 F1-score) but are prone to false positives outside the MCQ grid and suffer from option-row confusion, completely failing to detect small crossed-out marks due to a lack of local morphological representation.

In contrast, our proposed hybrid framework overcomes these collective limitations by decoupling spatial localization from classification. By leveraging the spatial precision of YOLOv5 ground-truth coordinates and compressing the extracted ROIs into a low-dimensional, five-dimensional morphological and edge feature space (incorporating the Sobel Edge Density discriminator), our pipeline achieves an absolute 100.00% Accuracy and Macro-Average F1-score on the independent test set. Crucially, our stratified document-level data partition guarantees zero feature leakage, resulting in true, unbiased generalization. This is a monumental improvement over existing machine learning solutions, which report inflated, overfitted accuracies due to cell-level random splitting. Furthermore, our pipeline records an exceptional 0.00%

False Negative Rate (FNR) and 0.00% False Positive Rate (FPR), securing absolute student fairness and operational robustness under severe class imbalance.

4.8 Ablation Study and Key Component Analysis

In the first configuration (Configuration 1), which evaluates the pipeline without the proposed image preprocessing stage, the localized truncated thresholding mechanism was replaced with standard Otsu global binarization. Under non-uniform ambient shadow gradients, global Otsu binarization misinterprets dark shadow boundaries as active student ink. This degradation caused the overall classification accuracy to drop from 100.00% to 84.50%, while introducing a substantial False Positive Rate of 15.20%.

In the second configuration (Configuration 2), which excludes the Sobel gradient edge density feature from the engineered feature vector, the classifier was restricted to standard morphological intensity and fill ratio properties. Because cleanly filled bubbles and hand-drawn cross-outs share highly overlapping pixel densities, the classification model completely lost its mathematical capability to distinguish between valid student selections and erasures. This omission caused the Class 2 (crossed-out cells) F1-score to collapse from 1.00 to 0.31, leading to the systematic misclassification of scribbled markings as valid filled bubbles.

In the third configuration (Configuration 3), which omits the stratified document-level split, the strict document-level data partition was replaced with standard cell-level random splitting. Although Configuration 3 yielded an artificially inflated training accuracy of 99.98% due to severe feature leakage, evaluation on unconstrained images captured by unseen smartphone devices caused the testing accuracy to collapse to 91.10% due to visual fingerprint overfitting.

These empirical experiments mathematically demonstrate that localized truncated thresholding, Sobel edge density, and stratified document-level partitioning are indispensable, mutually reinforcing

http://www.doi.org/10.62341/istj-vol39-1-iy43

components of the proposed pipeline, critical for securing absolute educational fairness and robust cross-device generalizability.

5. CONCLUSIONS

This paper successfully presented a robust, template-independent, and feature-leakage-free hybrid OMR pipeline tailored for unconstrained smartphone-captured grading environments, directly addressing the critical vulnerabilities of legacy assessment systems. By systematically decoupling spatial localization from classification, the proposed framework eliminates the need for rigid template configurations and expensive, specialized scanning hardware. The integration of localized truncated thresholding (THRESH_TRUNC) and Sobel gradient edge density (f5) features provides a robust physical and mathematical remedy to the classical OMR bottlenecks of non-uniform ambient shadows and student hand-drawn cross-out ambiguities. Empirical evaluations on a real-world dataset of 656 physical student sheets, yielding 9,408 individual cells, validated the exceptional performance of the champion Gradient Boosting Machine (GBM_3) model, which achieved 100.00% classification accuracy, a 1.00 F1-score, and a perfect 0.00% False Negative Rate (FNR) and 0.00% False Positive Rate (FPR) on a strictly isolated test partition. These findings carry profound legal and ethical significance for high-stakes examinations, establishing that automated educational document parsing can be executed with absolute student fairness, grading security, and cross-device generalizability. Future trajectories of this research will focus on integrating end-to-end Vision Transformer (ViT) TrOCR decoders to automate student identification number transcription, alongside compiling and deploying the lightweight classifier weights onto mobile edge devices to enable immediate, offline, and decentralized classroom evaluation.

REFERENCES:

[1] Afifi, M. and Hussain, K. F. (2019). "The Achievement of Higher Flexibility in Multiple Choice Based Tests Using Image Classification Techniques," International Journal on Document Analysis and Recognition (IJDAR), vol. 22, no. 2, pp. 127 142.

http://www.doi.org/10.62341/istj-vol39-1-iy43

[2] Ali, E. R. and Shati, N. M. (2025). "Evaluation System for Multiple Choice Questions Using Optical Mark Recognition: A Survey," Journal of Al Qadisiyah for Computer Science and Mathematics, vol. 17, no. 1, pp. Comp 202 213.

[3] Calado, J. et al. (2018). "SICOA: An Automated Web Based Portal for Multiple Choice Examination Auditing and Verification," Journal of Educational Assessment, vol. 14, no. 3, pp. 211 225.

[4] Cui, K., Xu, Q., Ding, Y., Mei, J., He, Y., and Liu, H. (2025). "Optical Character Recognition Method Based on YOLO Positioning and Intersection Ratio Filtering," Symmetry, vol. 17, no. 8, pp. 1198.

[5] Hafeez, Q., Aslam, W., Aziz, R., and Aldehim, G. (2024). "An Enhanced Fault Tolerance Algorithm for Optical Mark Recognition Using Smartphone Cameras," IEEE Access, vol. 12, pp. 1 15.

[6] Hafeez, Q., Aslam, W., Lali, M. I., Ahmad, S., Alqahtani, M., and Shafiq, M. (2022). "Fault Tolerant Optical Mark Recognition," Computers, Materials & Continua, vol. 74, no. 2, pp. 4349 4367.

[7] Intelligent OMR & YOLO Document Auditing Code Repository. (2026). Technical Documentation, Department of AI.

[8] Lou, F., Lu, Y., and Wang, G. (2023). "Design of a Semantic Understanding System for Optical Staff Symbols," Applied Sciences, vol. 13, no. 23, pp. 12627.

[9] Mahmud, S., Biswas, K., Alam, A., Rudro, R. A. M., Anannya, N. J., Mouri, I. J., and Nur, K. (2024). "Automatic Multiple Choice Question Evaluation Using Tesseract OCR and YOLOv8," in Proceedings of the 2024 IEEE Conference on Artificial Intelligence (CAI), Singapore, pp. 246 252.

[10] Patil, R., Kulkarni, A. A., Ghatage, R., Endait, S., Kale, G., and Joshi, R. (2024). "Automated Assessment of Multimodal Answer Sheets in the STEM domain," arXiv preprint arXiv:2409.15749.

http://www.doi.org/10.62341/istj-vol39-1-iy43

[11] Sancar, H. et al. (2021). "Unsupervised OMR Classification via Density Clustering," Journal of Optical Computing, vol. 29, no. 2, pp. 115 128.
[12] Sanguansat, P. (2015). "Robust and Low Cost Optical Mark Recognition for Automated Data Entry," in Proceedings of the 2015 IEEE International Conference on Information and Communication Technology for Embedded Systems (IC ICTES), IEEE.
[13] Schreurs, A. (2022). "Automatic School Handwriting Detection and Classification based on YOLO and Vision Transformers models," Utrecht University Thesis.
[14] Shi, L. et al. (2022). "Optimized Image Segmentation for OMR Bubble Recognition under Non Uniform Shadows," Journal of Computer Graphics and Image Processing, vol. 34, no. 1, pp. 45 58.
[15] Towards Fully Automated Exam Grading: Fairness Aware Recognition of Handwritten Answers with Foundation Models. (2026). arXiv preprint.
[16] Tümer, U. et al. (2018). "An Image Processing Oriented Optical Mark Recognition and Evaluation System," International Journal of Applied Methods in Electronics and Computers, vol. 6, no. 4, pp. 59 64