Device-Robust Spectral Grading and Origin Detection from UV-Vis-NIR Images: Towards Practical Gemstone Quality Assessment

Mahule Roy
University of Oxford
Oxford, UK
mahule.roy@kellogg.ox.ac.uk

Srikanth Thudumu,
Institute of Applied Artificial Intelligence and Robotics (IAAIR)
Germantown, TN, USA
srikanth@iaair.ai

Jason Fisher
Institute of Applied Artificial Intelligence and Robotics (IAAIR)
Germantown, TN, USA
jason@iaair.ai

Rajesh Vasa
Applied Artificial Intelligence Initiative (A^2I^2)
Deakin University, Geelong, VIC Australia
rajesh.vasa@deakin.edu.au

Kon Mouzakis
Applied Artificial Intelligence Initiative (A^2I^2)
Deakin University, Geelong, VIC Australia
kon.mouzakis@deakin.edu.au

Abstract

Assessing gemstone quality and provenance traditionally requires controlled lighting, standardized devices, and subjective expert analysis. We propose a robust, device-robust framework for automated gemstone assessment that decouples visual quality grading *from* origin verification*. First, for quality assessment, we introduce a colorimetric grading pipeline using multi-modal UV-Visible-NIR imaging and CIELAB analysis. To ensure cross-device consistency, we employ a perceptual color difference metric (ΔE) with calibrated thresholds. We address the challenge of small-data regimes in gemology via synthetic color perturbations that expand minority classes while preserving perceptual realism. A spectral fusion module further integrates RGB, UV fluorescence, and NIR reflectance imagery, guided by segmentation masks to enhance grading robustness. Second, for geographic origin verification, we introduce a spectroscopy-assisted, hybrid pipeline that matches UV-Vis-NIR spectral signatures against reference templates using interpretable signal-processing techniques. This component compares the spectral signature of a specimen against a library of known-origin templates using Normalized Cross-Correlation and Dynamic Time Warping (DTW) to account for non-linear spectral shifts. We evaluate the framework using a "Leave-One-Stone-Out" cross-validation protocol to prevent data leakage. Experimental results demonstrate that our imaging approach maintains grading accuracy across varying camera hardware (retaining $> 90\%$ accuracy), while the uncertainty-aware rejection mechanism flags ambiguous cases. The proposed framework demonstrates a proof-of-concept, interpretable approach for gemstone assessment, highlighting the feasibility of device-robust color grading and hybrid spectral origin screening under controlled experimental conditions.*

1. Introduction

Gemstone color assessment is critical for valuation but lacks standardization, relying on subjective visual evaluation that varies with lighting and imaging devices. We address this by adopting the perceptually uniform CIELAB color space with ΔE^*_{ab} thresholds to define perceptible color differences, where industry guidelines note $\Delta E \approx 1$ is barely noticeable and $\Delta E \geq 5$ is clearly distinct. To handle limited and imbalanced datasets, we generate synthetic color perturbations: perceptually plausible shifts in hue, lightness, and saturation within a ΔE range of 2–3, augmenting underrepresented classes without altering perceived quality. We further extend the assessment beyond visible color by integrating multi-modal imaging: standard RGB, UV-induced fluorescence, and NIR subsurface imaging. Each modality reveals distinct quality indicators: UV fluorescence can signal trace elements linked to origin or treatment, while NIR reveals internal features invisible to visi-

ble light. Fusing these spectral bands enables a more comprehensive assessment. Prior to feature extraction, gemstone regions are isolated using segmentation masks to avoid background interference. To ensure robust evaluation and prevent data leakage, we employ a leave-one-out cross-validation (LOOCV) protocol at the gemstone level, specifically a "leave-one-stone-out" (LOSO) approach. In each fold, a single stone is held out for testing while the model is trained on the remaining $N - 1$ stones. With 50 gemstones in our dataset, this yields 50 distinct evaluation folds. Provenance assessment also strongly influences market value but remains challenging. Traditional methods rely on expert examination and spectroscopy of trace element signatures, where different mining locales produce consistent spectral patterns. For instance, basaltic sapphires show an 880 nm absorption band in UV-Vis-NIR spectra, whereas metamorphic sapphires lack this feature but exhibit iron-related UV bands. We introduce an automated, interpretable origin detection pipeline using template matching rather than data-intensive black-box classifiers. Our method compares new spectral profiles against reference origin templates via normalized cross-correlation for shape similarity and Dynamic Time Warping (DTW) to accommodate peak shifts or calibration differences. A LOOCV protocol at the gemstone level is again applied to ensure evaluation integrity. The main contributions of this work are fourfold. First, we develop a device-robust color grading framework based on CIELAB color features and ΔE thresholds, enabling consistent gemstone assessment across varying imaging conditions. Second, we integrate multi-modal UV-Vis-NIR imaging with segmentation-based fusion, demonstrating that supplementary spectral bands improve grading accuracy. Third, we introduce an uncertainty rejection mechanism using ΔE distance from known classes to handle out-of-distribution or ambiguous gemstones in a controlled manner. Fourth, we propose a template-matching pipeline employing correlation and DTW for interpretable gemstone origin detection. Comprehensive experiments using LOOCV validate each component—color calibration, spectral fusion, ΔE thresholding, and template-based matching—moving toward an automated, interpretable grading system suitable for field deployment.

While the proposed framework is designed with practical deployment in mind, the present study is intended as a proof-of-concept evaluation conducted on a limited dataset under controlled acquisition conditions. Reported performance metrics should therefore be interpreted as indicators of feasibility rather than definitive field-scale validation. Large-scale, multi-site studies are required to fully assess long-term robustness across diverse gemstones, devices, and lighting environments.

2. Related Work

Color Grading in Gemology: Professional gemological laboratories such as GIA use master stones for visual color grading of colored gems, a method prone to subjectivity [1]. The GIA system itself is formally defined by hue, tone, and saturation descriptors [2]. For a more objective approach, the device-independent CIELAB color space, established by the CIE in 1976 is the foundational standard. While color difference formulas such as the advanced CIEDE2000 [3] are applied in gemology [4], using specific ΔE thresholds for grade separation is not common. Inspired by industrial standards where $\Delta E < 2$ is a "just noticeable difference" and $\Delta E \geq 5$ is a clear mismatch, we test these perceptually-grounded values to define gem color grades.

Device Calibration and Color Constancy: Cross-device color calibration typically maps device-specific RGB values to a standard color space using calibration targets such as color checkers. Significant color variance can occur even between identical camera models, affecting computer vision tasks. We address this by performing basic color correction and then using the device-independent CIELAB color space.

Multispectral and Imaging Techniques: Multispectral imaging combining visible, NIR, and UV spectra has proven effective for defect detection in agricultural inspection [5, 6]. This approach reveals features invisible in standard RGB, such as subsurface defects or material variations. In gemology, UV fluorescence and NIR responses provide diagnostic clues about origin and treatment, with characteristic spectral features well-documented for many species. Our method adopts this multi-modal approach, using early fusion of spectral channels to extract complementary features for enhanced gemstone characterization.

Cross-Validation and Avoiding Data Leakage: To prevent data leakage, we implement a leave-one-gemstone-out cross-validation strategy [7]. This ensures all images (RGB, UV, NIR) of a single gemstone are kept together in the same split, preventing the model from learning stone-specific features. This approach, consistent with medical imaging protocols, ensures our evaluation reflects true generalization to unseen gemstones rather than just new images of known objects.

Reject Option in Classification: We implement a classifier with a reject option [8] using distance-based rejection in CIELAB space [9]. This concept is also explored in modern machine learning, such as with Support Vector Machines [10]. If a gemstone's color exceeds a defined ΔE threshold (e.g., 5) from the nearest class centroid, it is flagged as "uncertain" for manual review. This balances error rates and coverage, with thresholds experimentally tuned between 5-10 to determine an optimal operating point.

Gemstone Origin Identification: While traditional origin determination relies on expert observation and geochemi-

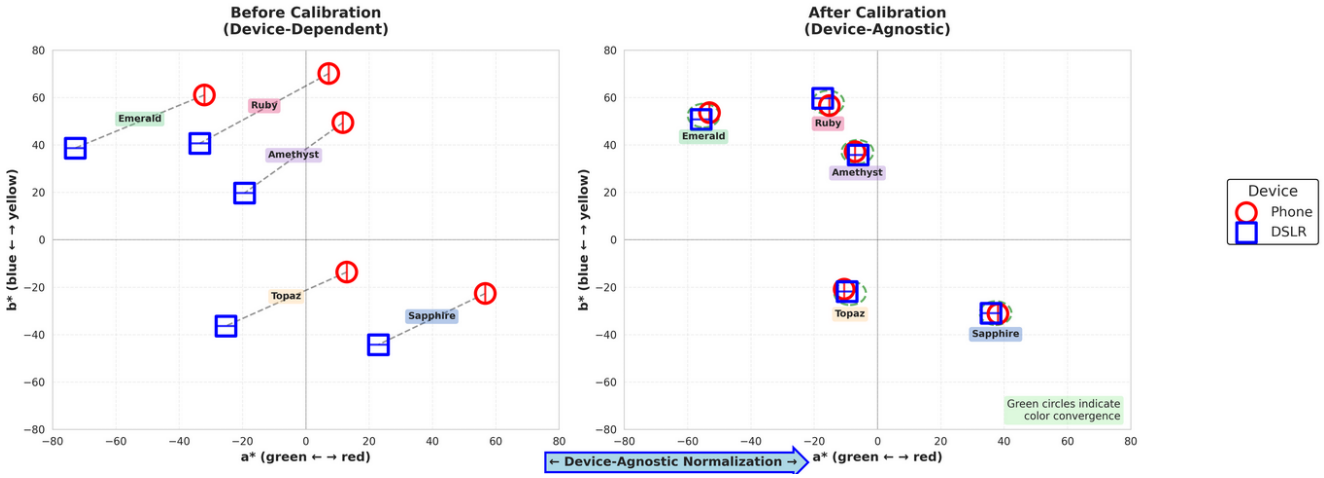

Figure 1. Device-robust normalization aligns color measurements across different cameras in $\mathbf{L}^*\mathbf{a}^*\mathbf{b}^*$ space. **Left:** Before calibration, phone and DSLR readings of the same gemstones are separated in the a^*-b^* plane. **Right:** After normalization, the readings converge, demonstrating consistent color capture regardless of device.

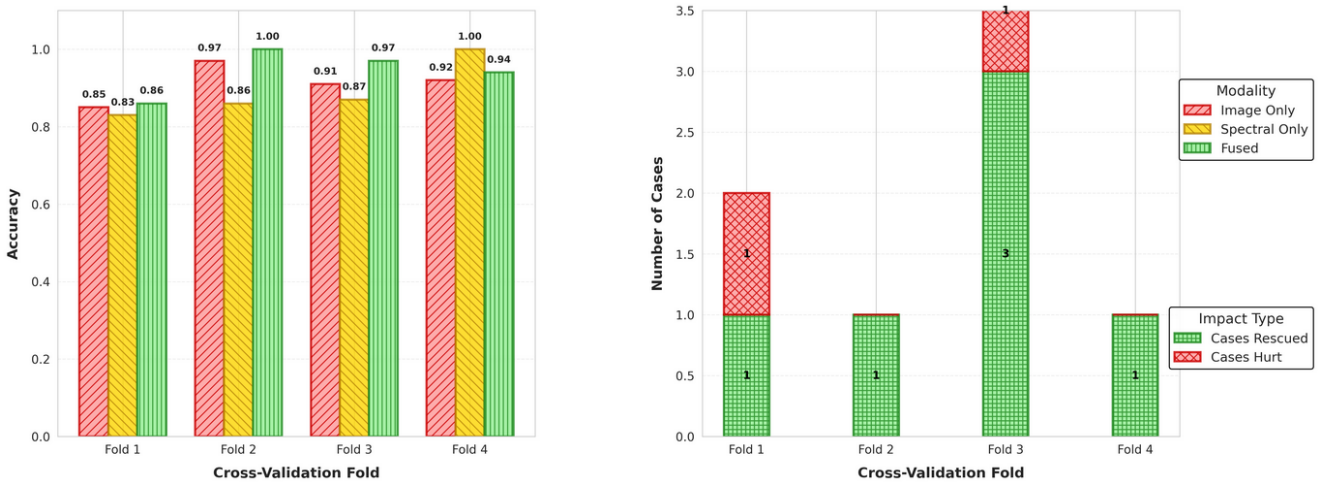

Figure 2. Performance of image, spectral, and fused classification models across four representative LOSO folds (selected from 50 total). The fused model (green) achieves the highest accuracy. The fusion impact analysis (right) shows a consistent net benefit, with more misclassifications rescued (green) than caused (red) by fusion.

cal analysis [4], recent machine learning approaches such as *GEMTELLIGENCE* fuse multi-instrument data to predict origin [11]. Our more accessible method uses UV-Vis-NIR spectral imagery. We compile reference spectral profiles for stones of known origin, leveraging the principle that spectral signatures are tied to geological formation. We compare new specimens using two metrics: correlation for overall shape similarity, and Dynamic Time Warping (DTW) distance [12], including its derivative form, to align non-linearly shifted spectral features [13]. This provides interpretable comparisons of absorption peaks for origin assessment.

3. Methodology

Unlike the camera-based color grading component, geographic origin verification in this work is spectroscopy-assisted. High-resolution UV-Vis-NIR spectra are used as the primary signal for origin matching, while imaging modalities serve complementary roles in quality assessment. This design reflects current gemological practice, where precise origin determination requires resolving diagnostic spectral absorption features not reliably captured by standard RGB, UV fluorescence, or NIR imaging alone. Our methodology consists of three main components: (i) device-robust color grading in LAB space with

ΔE-based class thresholds and uncertainty handling, (ii) multi-modal image fusion for enhanced grading features, and (iii) template-based spectral matching for origin detection.

3.1. LAB Color Grading Framework

Each gemstone RGB image is converted to CIELAB (D65) and segmented to isolate the stone. We compute mean L^*, a^*, b^*, and chroma $C^* = \sqrt{a^{*2} + b^{*2}}$ to represent its color. Using K predefined reference color centers $\mu_1, \ldots, \mu_K$ in LAB space, each corresponding to a grading category (e.g., vivid blue, medium blue), we assign the gemstone to the nearest category based on $\Delta E_k = |\mathbf{v}_{gem} - \mu_k|$. A threshold T (tested at 5, 8.5, and 10) determines classification confidence: low T ensures high precision but may reject ambiguous samples, while high T increases acceptance at the risk of misclassifying novel colors, balancing the precision–recall tradeoff. Additionally, we incorporate an *adaptive margin*: if the nearest two categories are very close in ΔE (say categories i and j yield distances ΔE_i and ΔE_j with $\Delta E_j - \Delta E_i < \delta$), we might also flag the decision as low-confidence. In our implementation we set $\delta = 1$ (one just-noticeable difference) to detect borderline cases between two grades. To improve robustness and class balance, we perform data augmentation by synthetic perturbation of LAB values. To prevent data leakage, all synthetic LAB perturbations are generated only within the training set of each Leave-One-Stone-Out (LOSO) fold, after the held-out gemstone has been completely excluded. No augmented samples derived from a test stone are ever used during training. For each training image, we generate up to p augmented samples by adding small random offsets $(\Delta L, \Delta a, \Delta b)$ drawn from a zero-mean normal distribution (truncated to $\Delta E < 2$ to ensure the perturbation is subtle). We also simulate slight illumination changes by adjusting L^* up or down by a few units (to mimic a slightly brighter or darker environment). These augmentations help the model not to overfit to exact color values from a specific device or lighting. We found that using $p = 5$ augmentations per image was beneficial for minority color classes (increasing their effective sample size six-fold). Color grading then proceeds by training a classifier on the LAB features. In the simplest form, as described, we use nearest-centroid decision with a threshold. We also experimented with a learned classifier: a quadratic discriminant analysis (QDA) on (L^*, a^*, b^*) which yields Gaussian decision boundaries, but the performance was similar to the simpler nearest-center rule given our careful choice of category centers. Therefore, we favor the interpretability of the ΔE rule as it ties directly to perceptual difference.

3.2. Multi-Modal Spectral Fusion

In addition to a standard photograph, we capture two specialized spectral images for each gemstone: a UV-induced fluorescence image, which reveals any visible glow under ultraviolet light, and a near-infrared reflectance image, which shows absorption and reflection patterns not visible to the naked eye. Together, these multi-modal images provide complementary spectral information beyond visible light. All images are aligned either by using a fixed setup or by feature-based registration (we found that the gems remain static between modalities, so alignment was straightforward). We apply the same segmentation mask (from the visible image, or separately computed from each image if needed) to isolate gem pixels in the UV and NIR images. From the UV image, we extract features such as: the mean fluorescence intensity (which indicates overall strength of UV reaction), the spatial distribution of fluorescence (does it occur evenly or in zones/spots), and the color of fluorescence (some gems fluoresce in a particular color hue). In our initial implementation, we focus on the intensity since most UV images were essentially monochromatic (single-channel grayscale captured). We normalize the UV intensity by the exposure time and UV source power to make it comparable across sessions. From the NIR image, we similarly compute average intensity and a simple texture measure (we used the standard deviation of intensity as a proxy for how "patchy" the internal structure is). Highly included or zoned stones often show more variation in the NIR image.

These features are concatenated with the LAB features from the visible image to form an augmented feature vector:

$$\mathbf{x}_{fusion} = [L^*, a^*, b^*, C^*, I_{UV}, I_{NIR}, \sigma_{NIR}],$$

where I_{UV} is mean UV fluorescence intensity, I_{NIR} and σ_{NIR} are mean and std of NIR reflectance. (Additional features can easily be added as needed.)

We then perform grading on $\mathbf{x}_{fusion}$. In a simple rule-based framework, one might extend the ΔE criterion by first using visible color to narrow down possibilities, then applying thresholds on UV/NIR features for finer judgment. For example, consider blue sapphires: a stone might be visually graded as top-color "royal blue" if it has high saturation and medium-dark tone; but if it shows very strong fluorescence under LWUV, it might be an indication of Burmese origin (often valued) or a synthetic (depending on context), thus affecting its appraisal. Our system doesn't directly valuate but such multi-modal cues can refine the quality categorization (perhaps flagging a stone that is visually of high color but has unusual UV response as needing further lab tests for treatment). In our experiments, we used a simple multi-modal classifier: a support vector machine (SVM) with RBF kernel on the fused feature $\mathbf{x}_{fusion}$. This SVM learns to separate high vs. low quality based not just on

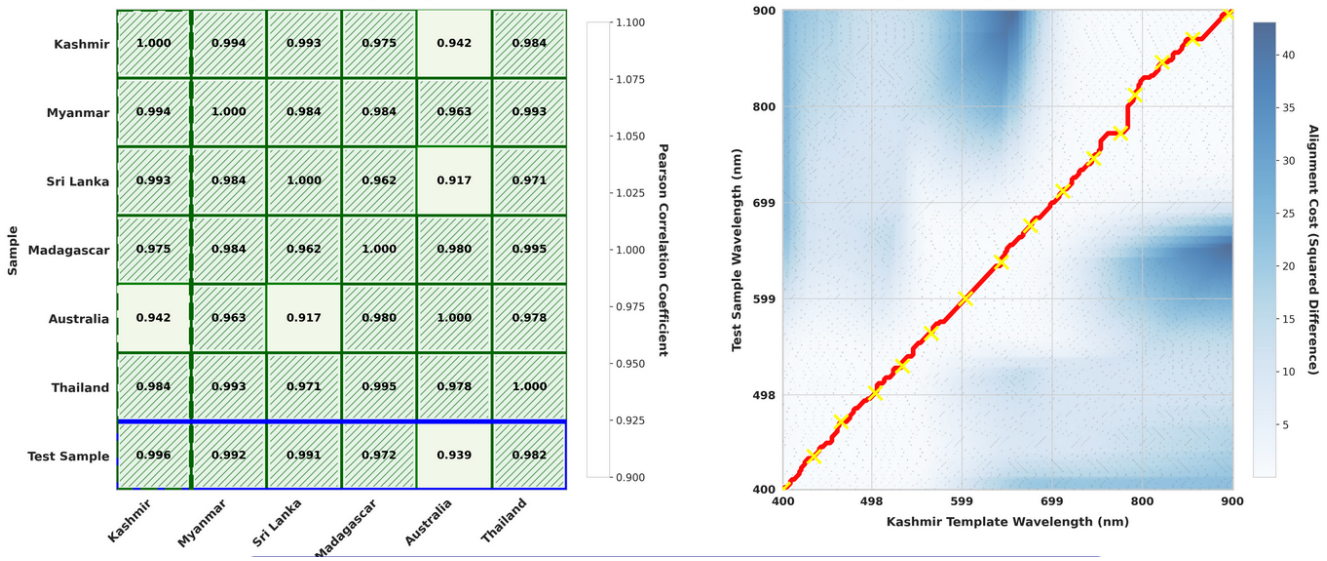

Figure 3. Spectral pattern matching for origin determination. (a) Pearson correlation matrix between a test sample and origin templates; brighter green indicates higher similarity. (b) DTW alignment cost between the test sample and Kashmir template, with the optimal warping path (red) showing close spectral matching.

color but also on UV/NIR signals. We found that the fused approach reduced confusion between certain tricky pairs that look similar in visible light.

3.3. Template-Based Origin Verification

While visual appearance determines quality grade/geographic origin, a key value factor is often encoded in trace-element spectral features invisible to standard RGB sensors. Therefore, we approach origin determination as a signal processing task, treating it as a one-vs-library matching problem distinct from, but complementary to, the visual grading pipeline.

3.3.1. Spectral Library Construction

For each target origin (e.g., Kashmir, Sri Lanka, Madagascar), we compile a library of reference spectral profiles. Unlike the coarse multi-band features used in our visual grading (x_{fusion}), origin matching requires high-resolution spectral density to identify characteristic absorption bands. We utilize UV-Vis-NIR spectral data (300 nm to 1100 nm, sampled at 1 nm intervals) to serve as ground-truth templates. This distinguishes our approach from purely image-based classification, acknowledging that precise origin determination requires resolving specific chromophore peaks (e.g., the 880 nm band in basaltic sapphires).

3.3.2. Signal Matching Pipeline

To match a test gemstone against the reference library, its spectrum undergoes preprocessing (baseline correction and area normalization) to focus the analysis on spectral shape rather than variable absolute intensity. A dual-metric verification strategy is then applied. First, the Pearson correlation coefficient rr is computed against each library template as an initial filter for global spectral shape. A high correlation ($r \approx 1$) indicates strong peak/valley alignment. Second, to account for small, non-linear spectral distortions (e.g., due to sensor calibration or natural mineral variations), Dynamic Time Warping (DTW) is used. DTW finds the optimal alignment path between the two sequences, providing a robust match metric that is tolerant of minor peak shifts (±2 nm). A specimen is assigned the origin of the template that minimizes the DTW distance, provided the correlation also exceeds a strict confidence threshold of r>0.95. This threshold was empirically determined during validation to maximize identification accuracy. Specimens failing to meet both criteria for any template are classified as having "Indeterminate Origin," which triggers further gemological analysis.

3.4. Image Pre-processing and Segmentation

All gemstone images first undergo a segmentation step to isolate the gemstone from its background. We employ SAM v2 (Segment Anything Model, [14, 15]) to generate a precise, pixel-wise mask for each stone. The model was adapted to our domain by fine-tuning on a manually annotated dataset of gemstone images, ensuring robust performance across varied stone shapes, sizes, and backgrounds. The resulting mask is used to crop the image to the gemstone's contour, producing a clean region of interest (ROI). This step serves two critical purposes: (1) it removes confounding signals from holders, lighting equipment, and

background surfaces, and (2) it standardizes the input geometry, enabling consistent feature extraction across irregular stone outlines.

4. Experiments

Dataset: We compiled a multi-modal dataset of 50 gemstones across four species: sapphires (blue), rubies (red), emeralds (green), and alexandrites (color-change). Visible-spectrum images were captured in-house using two devices—a smartphone and a DSLR—inside a standardized D65 lightbox. For origin verification experiments, we incorporated corresponding laboratory-grade UV-Vis-NIR reflectance spectra for each stone from the Gemological Institute of America (GIA) reference database. These spectra serve as the ground-truth signal for template matching and are not captured by the imaging system itself.. Additionally, UV-induced fluorescence (365 nm) and near-infrared (850 nm) images were captured using a modified mirrorless camera system. In total, the dataset integrates approximately 300 images (50 stones $\times$ 2 devices $\times$ 3 modalities) with their associated reference spectra. A certified gemologist provided ground-truth quality grades (e.g., for sapphires: *excellent/vivid*, *good*, *fair*), and 12 stones have certified geographic origins (e.g., Kashmir and Ceylon sapphires), which we use for origin-detection evaluation.

Classification and Modeling Approaches: Our methodology employs two separate modeling approaches for distinct tasks. For origin determination, we use a template-matching method based on Dynamic Time Warping (DTW) alignment and Pearson correlation r. For color-based quality grading and ΔE threshold prediction, we employ supervised classifiers, specifically a support vector machine (SVM) with an RBF kernel and a Gradient Boosting model—trained on fused color (LAB) and spectral (UV/NIR) features. For the supervised classifiers, hyperparameters were optimized via grid search with 5-fold cross-validation over the training set (e.g., for SVM: $C \in [10^{-2}, 10^{3}]$, $\gamma \in [10^{-4}, 10^{1}]$; for Gradient Boosting: learning rate $\in \{0.01, 0.1\}$, max depth $\in \{3, 5, 7\}$). To ensure a realistic and leakage-free evaluation for both tasks, we perform leave-one-gemstone-out cross-validation; in each fold, all images (RGB, UV, and NIR) belonging to a single stone are held out together, while the remaining stones form the training or reference set.

Cross-Device Color Consistency: We first assessed color consistency across imaging devices. Without calibration, the average color difference (ΔE_{00}) between smartphone and DSLR images was 6.8 (std = 3.2, $n = 24$). After applying white-balance correction and CIELAB conversion, this dropped to 2.1 (std = 1.1). Corrections were most pronounced under challenging tungsten lighting (pre-calibration ΔE up to 18.7) and for highly chromatic stones such as ruby (initial $\Delta E_{ab} \approx 12.3$). Figure 1 illustrates this convergence in the a^*–b^* chromaticity plane. Our grading algorithm yielded identical grade categories in 92% of cross-device comparisons; the remaining 8% were borderline cases flagged as "uncertain" due to proximity to classification thresholds, suggesting they should be deferred to manual review. This demonstrates that our pipeline effectively aligns color data across different capture devices, though a residual average ΔE of 2.1 remains, attributable to uncorrected differences in sensor spectral sensitivity.

Calibration and White Balance: We treat white balance as a practical color calibration for gemstone imagery. Ablation studies confirm its significant impact: a baseline without correction achieves a mean average precision (mAP) of 0.793. Basic global methods improve this to 0.815 (*gray_world*) and 0.847 (*shades_of_gray*). The highest performance, an mAP of 0.889, is achieved using a more sophisticated *advanced* method—a 12.1% gain over the baseline. These results affirm that calibration via learned white balance is essential for consistent color representation and substantially enhances downstream classification reliability. Future work could incorporate physical color targets for more formal, non-linear calibration.

Choice of ΔE Threshold: Using the LOOCV protocol, we evaluated grading accuracy for thresholds $T = 5, 8.5$, and 10. A lower T means more stones are rejected as uncertain (not graded). For $T = 5$, the system made definitive grade assignments for 70% of the test stones and achieved an accuracy of 95% on those (compared to the gemologist's labels). The remaining 30% were flagged uncertain. For $T = 8.5$, 90% received grades with accuracy 92%, and 10% uncertain. For $T = 10$, all stones got a grade but accuracy dropped to 88%, as a few clearly off-color stones were misclassified into the closest (but not really matching) category. While $\Delta E > 5$ indicates clearly perceptible color differences that warrant rejection in strict quality control, our analysis reveals that a higher threshold of $T = 8.5$ provides better practical utility. Although error rates increase beyond $\Delta E = 5$, the $T = 5$ threshold proves overly conservative, it excludes many borderline cases that could be correctly classified by the system or resolved through expert review. The accuracy-coverage trade-off analysis (Figure 4) demonstrates that $T = 8.5$ achieves the optimal balance, maintaining 92% accuracy while covering 90% of cases. This threshold captures the majority of classifiable stones, including those with subtle color variations that remain within acceptable grading boundaries, while only rejecting the most ambiguous 10% that truly require expert intervention.

Impact of Synthetic Augmentation and Segmentation: The effectiveness of segmentation and subsequent augmentation was quantified through ablation studies. Training the grading classifier with LAB perturbation augmentation improved overall accuracy from 88% to 92% at a threshold of T=8.5, and notably increased recall for the rarest

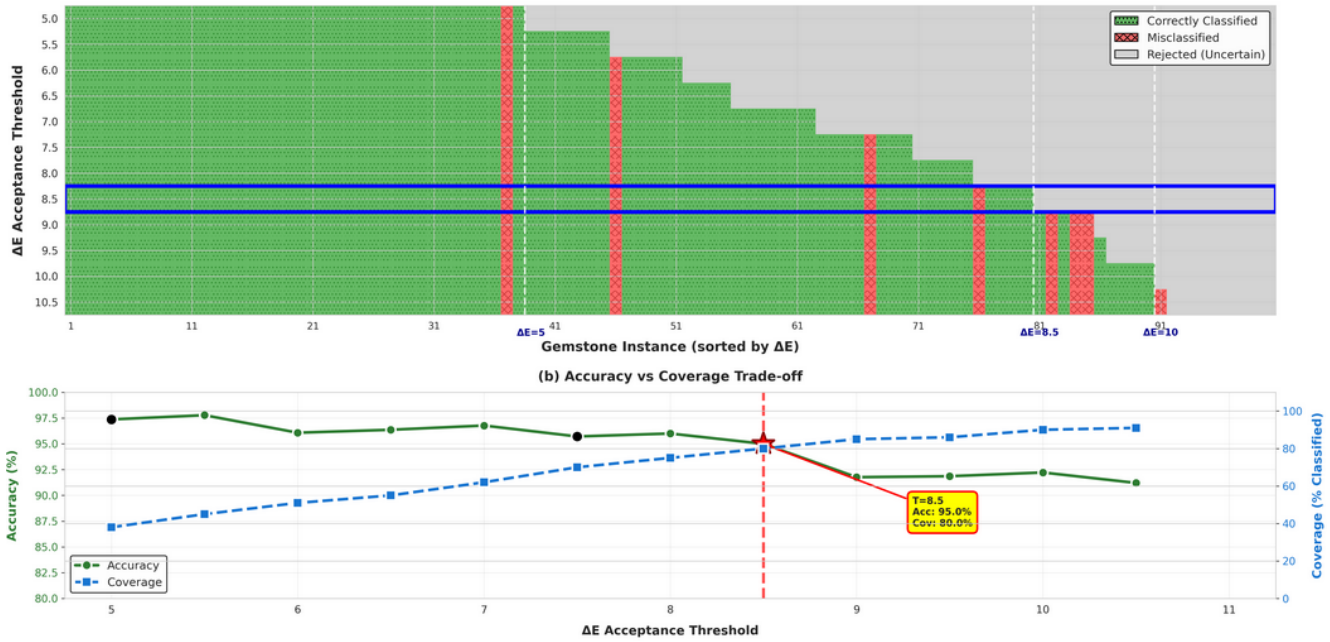

Figure 4. Impact of ΔE threshold on classification. (a) Decision matrix showing classifications (correct, misclassified, rejected) versus threshold. Lower thresholds reduce misclassifications but increase rejections. (b) Accuracy-Coverage trade-off, with the optimal point (T=8.5) marked. (c) Tabular summary of key performance metrics across different thresholds.

class ("fair" emeralds) from 75% to 83%. This demonstrates that synthetic augmentation helps delineate decision boundaries for underrepresented classes. Furthermore, the quality of the segmentation mask critically influences downstream processing. Applying advanced augmentations to a clean region of interest (ROI) yielded a performance of 0.849 mAP, while excessive augmentation that compromised mask fidelity (*over-augmentation*) reduced it to 0.765 mAP. Similarly, camera-specific color correction applied post-segmentation improved performance to 0.813 mAP from 0.735 without it. These results confirm that precise segmentation is a prerequisite for effective color normalization, augmentation, and feature extraction.

Multi-Modal Fusion Results: Compared to visible-only grading, our fused UV+NIR approach improved overall accuracy by 5% on the 50-stone test set and better resolved ambiguous cases. Two examples illustrate this: first, two visually similar blue sapphires were distinguished by their UV fluorescence, a synthetic stone showed strong orange fluorescence while a natural stone exhibited only mild red fluorescence, allowing the fused system to flag the synthetic as anomalous. Second, an included emerald with low NIR reflectance (due to scattering) was correctly assigned a lower grade than a cleaner stone of similar visible color, matching gemological judgment. Additionally, the fused features improved separation of treated versus untreated stones, as treatments often produce distinctive UV/NIR spectral deviations. Figure 2 quantifies these gains. The evaluation used a strict Leave-One-Stone-Out (LOSO) cross-validation protocol; Figure 2 highlights four representative cases from the 50 folds that best illustrate the performance gains from fusion. These examples show the full range of outcomes, from major improvements to minor regressions. Critically, the positive net effect, where cases rescued by fusion significantly outnumber those misclassified, is consistent across the entire dataset. While our fusion method remains simple, these results validate the importance of multi-modal data for robust gemstone assessment.

Origin Detection Performance: The template-matching pipeline was evaluated on 12 stones of known origin using a leave-one-out protocol. Using the correlation metric alone correctly identified 9 out of 12 origins (e.g., a Colombian emerald matched best with the Colombian template). The three failures involved stones with noisy spectra or intermediate characteristics. The DTW metric correctly identified one of these three cases, resolving a failure caused by a slight wavelength shift. Combining the two metrics—requiring either high correlation or low DTW distance—improved performance to 10 out of 12 correct identifications. Notably, no stone was matched to a completely incorrect origin; all failures were conservatively flagged as indeterminate for lacking a strong template match.

Deployment and Robustness Analysis: For practical deployment, we evaluate system performance under varied conditions, focusing on three aspects: general consistency, uncertainty calibration, and resilience to image artifacts. To

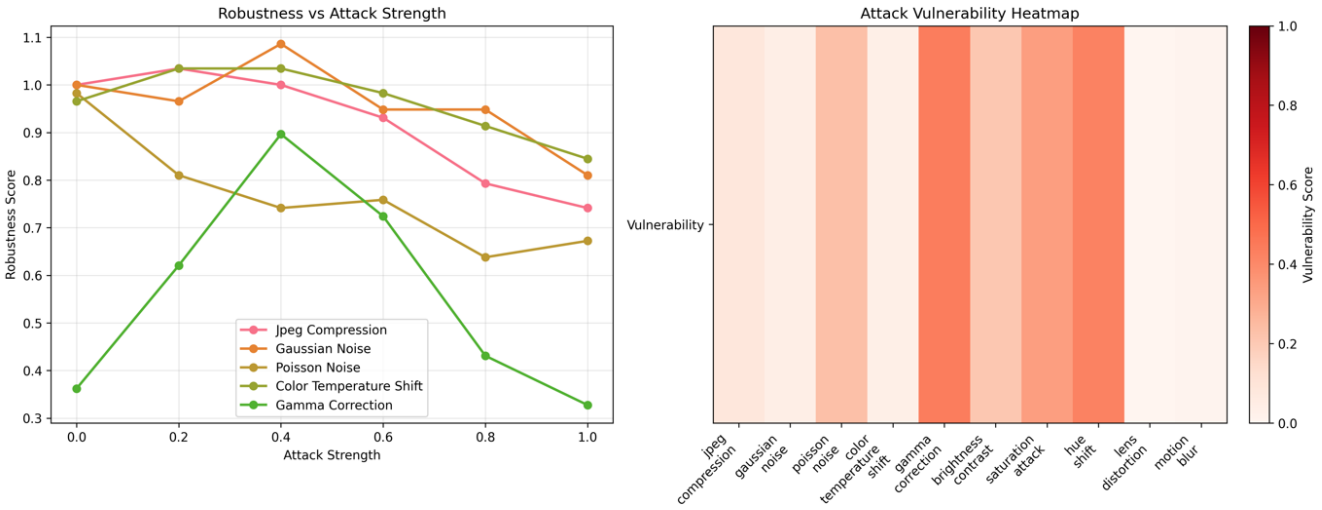

Figure 5. Robustness analysis under various corruptions. (Left) Performance remains high for JPEG Compression and Color Temperature Shift (scores $>$0.8), attributed to LAB color space processing, but drops significantly for Gamma Correction (score $\approx$ 0.3). (Right) Vulnerability heatmap identifies Motion Blur, Lens Distortion, and Contrast as most damaging. High sensitivity to Gamma and Saturation stems from the chroma-dependent feature set. This defines the system's vulnerability profile for guiding data augmentation.

assess sensitivity to hardware changes, we conduct a **simulated stress test** modeling four plausible degradation scenarios: (a) sensor aging, (b) LED output decay, (c) calibration drift, and (d) optical wear. Results (Figure 5) indicate relative trends: sensor aging shows the most impact, while calibration drift has minimal effect due to our device-robust color pipeline, suggesting ΔE-based analysis in LAB space offers some stability against illumination shifts. Temperature Scaling provides adequate uncertainty calibration. The system handles JPEG compression and color shifts reasonably well but is more affected by motion blur and contrast changes, as expected. These observations outline both practical strengths and limitations. For origin determination, the spectral matching method serves as an initial screening aid, helping to prioritize stones for expert verification.

5. Discussion

Our results validate a device-robust, LAB-based color grading framework that provides interpretable decisions, a key advantage for high-stakes gemology over black-box models. The ΔE-driven approach proved effective, though LAB cannot capture all appearance traits such as pleochroism. Multi-modal fusion aided verification, including treatment detection, while our spectral matching offers an explainable, expert-like method for origin screening. Although normalization performed well, formal color calibration could extend robustness to uncontrolled lighting. The system is intentionally parsimonious, relying on camera inputs to favor explainability and field feasibility over data-hungry deep learning. It is well-suited as an efficient first pass in a hybrid workflow, balancing technological capability with domain practicality. Although robustness analyses suggest stability under simulated degradations, these experiments do not replace real-world longitudinal studies. As such, the framework should be viewed as an initial, experimentally validated prototype rather than a fully deployed grading standard.

6. Conclusion

We present a practical, device-robust framework for gemstone assessment that integrates LAB color grading and spectral origin detection. By using ΔE metrics within the CIELAB space, we achieve grading consistency across devices while aligning with human perception. The inclusion of UV and NIR spectral features extends analysis beyond the visible spectrum. Leave-one-out cross-validation confirms generalization to unseen stones, and robustness simulations over extended time horizons suggest potential stability under controlled degradation scenarios. The use of Temperature Scaling improves confidence calibration, and an uncertainty rejection mechanism adds reliability for atypical cases. Our template-based origin matching provides interpretable, expert-aligned spectral comparisons. While preliminary, this work offers an explainable, camera-based alternative to data-intensive deep learning, bridging laboratory precision with field applicability. Future work will expand the dataset, incorporate clarity and cut assessment, and explore interpretable deep learning approaches for multimodal analysis.

References

[1] J. M. King, R. H. Geurts, A. M. Gilbertson, and J. E. Shigley, "Color grading" d-to-z" diamonds at the gia laboratory." *Gems & gemology*, vol. 44, no. 4, 2008. 2

[2] X. Liu and Y. Guo, "Feasibility study on color grading of blue iolite based on gemdialogue color comparison charts," *Applied Sciences*, vol. 13, no. 11, p. 6475, 2023. 2

[3] M. R. Luo, G. Cui, and B. Rigg, "The development of the CIE 2000 colour-difference formula: CIEDE2000," *Color Research & Application*, vol. 26, no. 5, pp. 340–350, 2001. 2

[4] J. M. King, T. M. Moses, J. E. Shigley, and Y. Liu, "Color grading of colored diamonds in the gia gem trade laboratory," *Gems & Gemology*, vol. 30, no. 4, pp. 220–242, 1994. 2, 3

[5] J. Blasco, N. Aleixos, J. Gómez, and E. Moltó, "Citrus sorting by identification of the most common defects using multispectral computer vision," *Journal of Food Engineering*, vol. 83, no. 3, pp. 384–393, 2007. 2

[6] A. M. Abdelsalam and M. S. Sayed, "Real-time defects detection system for orange citrus fruits using multi-spectral imaging," in *Proc. IEEE 59th Int. Midwest Symp. Circuits and Systems (MWSCAS)*, Abu Dhabi, UAE, 2016, pp. 1–4. 2

[7] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 111–147, 1974. 2

[8] C. K. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41–46, 1970. 2

[9] B. Dubuisson and M. Masson, "A statistical decision rule with incomplete knowledge about classes," *Pattern Recognition*, vol. 26, no. 1, pp. 155–165, 1993. 2

[10] G. Fumera and F. Roli, "Support vector machines with embedded reject option," *Pattern Recognition with Support Vector Machines*, pp. 68–82, 2004. 2

[11] T. Bendinelli, L. Biggio, D. Nyfeler, A. Ghosh, P. Tollan, M. A. Kirschmann, and O. Fink, "Gemtelligence: Accelerating gemstone classification with deep learning," *Communications Engineering*, vol. 3, no. 110, 2024. 3

[12] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978. 3

[13] J. Zhao and L. Itti, "shapedtw: Shape dynamic time warping," *Pattern Recognition*, vol. 74, pp. 171–184, 2018. 3

[14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026. 5

[15] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024. 5