VISHOD - Visual Dataset Scraping and Hybrid Outlier Detection

Flávio Moura[1], Vitor Melo[1], André Alves[1], Lyanh Pinto[1], Walter Júnior[1], Adriano Santos[1], Roberto Oliveira[2], Diego Cardoso[1], Marcos Seruffo[1]

[1] Institute of Technology – Federal University of Pará — Belém – PA – Brazil

[2]Institute of Exact and Natural Sciences – Federal University of Pará — Belém – PA – Brazil

flavio.moura@itec.ufpa.br, vitor.melo@itec.ufpa.br,

andre.neves.alves@itec.ufpa.br, lyanh.pinto@itec.ufpa.br,

walter@inteceleri.com.br, adriano.madureira.santos@itec.ufpa.br,

limao@ufpa.br, jmorais@ufpa.br, diego@ufpa.br, seruffo@ufpa.br

Abstract. *Supervised learning depends on data quality, yet manual labeling is costly for resource-limited environments. This paper presents an automated pipeline for building and refining image datasets via web scraping and hybrid outlier detection. The method integrates automated collection, deep learning semantic extraction, consensus-driven anomaly detection, and statistical validation. As a case study, a high-quality dataset without manual annotation was produced for the classification of 2D geometric shapes for educational use in the Amazon. Results show a 25.6% increase in average PCA variance, a 5.6% reduction in centroid distance, and a 21.6% gain in the structural similarity index, evidencing greater diversity, cohesion, and visual homogeneity in the data.*

1. Introduction

Image classification is a fundamental technique in areas such as robotics [Habuza et al. 2021] and Computer-Aided Design (CAD) [Kadhim et al. 2022]. It allows automated systems to identify and classify objects or patterns in images, facilitating real-time decision-making. In robotics, image classification helps robots understand their surrounding environment and execute actions based on visual recognition [Li et al. 2023]. In CAD, it enhances the precision and efficiency of design and production processes, optimizing everything from model creation to quality control [Xiao and Ni 2024]. The effectiveness of image classification models critically depends on the availability of extensive and properly labeled datasets, specific to the desired application [Gong et al. 2023].

However, obtaining large volumes of labeled data is complex, manual, and costly [Adnan et al. 2021]. This challenge, coupled with visual data growth, makes Automatic Image Annotation (AIA) crucial. Despite advances, AIA faces issues like high-dimensional features, semantic gaps, and data scarcity [Adnan et al. 2021]. Automatically collected data can be noisy, biased, or contaminated, compromising model effectiveness; and many methods lack systematic curation, cleaning, or outlier detection, impacting dataset quality and model performance [Guo et al. 2021]. Thus, integrated solutions

are needed to automate data collection, organization, and refinement for computationally efficient and semantically demanding tasks.

Web scraping effectively mitigates data scarcity by automating the creation of large, scalable, and thematic datasets essential for machine learning [Schuhmann et al. 2022]. However, raw scraped images often contain noise, semantic inconsistencies, and outliers that can impair model generalization and performance [Rodríguez-Rodríguez et al. 2024]. Consequently, robust filtering and preprocessing are critical to ensure dataset consistency, representativeness, and overall effectiveness in pattern recognition applications.

While data diversity can foster robustness, unmanaged variations in resolution, lighting, and context often compromise dataset uniformity and model generalization [Aghabagherloo et al. 2025]. This lack of standardization is particularly challenging in web scraping for niche applications like shape classification, where the absence of clear curation and labeling standards leads to disorganized datasets and reduced interoperability [Xu et al. 2024]. Such inconsistencies are especially detrimental to lightweight models, where noise directly impacts precision and stability [Sharma et al. 2022].

To address these challenges, this work proposes an automated, low-cost methodological pipeline for building image datasets through three integrated phases: (i) automated image collection via web scraping (Subsection 3.1); (ii) hybrid outlier detection and removal incorporating structural analysis, clustering, and anomaly detection (Subsection 3.3); and (iii) quality assessment based on visual similarity and statistical indicators (Subsection 3.2). This scalable approach ensures semantic consistency and robustness, providing a replicable model for high-quality data generation in resource-constrained environments.

The method's applicability was demonstrated through a case study developing an AI model for a geometry education mobile application called Geometa[1], designed to classify 2D shapes in offline settings and on low-performance devices. Due to the scarcity of domain-specific public datasets, the proposed automated pipeline was used to collect images for nine geometric classes via web scraping, followed by hybrid outlier filtering and statistical validation. This process enabled the development of a computationally efficient model, refined through hyperparameter optimization and supervised training, ensuring high data representativeness and seamless integration into the educational tool.

The main contributions are: (i) an automated method for building image datasets, integrating collection, cleaning, and quality assessment; (ii) a hybrid, multidimensional outlier detection approach combining perceptual and statistical metrics; (iii) demonstration of applicability in a real-world educational scenario; and (iv) a replicable, low-cost computational pipeline for classification tasks with scarce labeled data. The article is organized as follows: Section 2 presents related works; Section 3 describes the methodology; Section 4 details the case study; Section 5 discusses results; and Section 6 provides conclusions and future work.

[1]https://play.google.com/store/apps/details?id=com.Inteceleri.Geometa

2. Related Works

Several studies have addressed outlier detection and data acquisition through diverse methodologies. Tan et al. (2021) utilized Poisson interpolation for pixel-level anomaly detection, while Nassif et al. (2021) provided a systematic review of machine learning models for identifying anomalous data across domains such as cybersecurity and finance. Regarding data collection, Thota et al. (2021) demonstrated the efficacy of web scraping for building large-scale datasets, though focused on text-based sentiment analysis. Furthermore, Stewart et al. (2022) optimized the HDBSCAN* algorithm for advanced clustering and outlier identification, emphasizing the role of Euclidean distance in maintaining proximity relationships within complex data structures.

Despite these contributions, existing literature often treats collection and refinement as isolated processes or focuses on specific data types that do not align with visual classification needs. For instance, pixel-based detection lacks the semantic depth provided by neural feature vectors [Tan et al. 2021], and broad reviews typically overlook integrated pipelines for data curation [Nassif et al. 2021]. Similarly, while web scraping is proven for text, its application to image datasets requires specialized filtering to handle noise [Thota and Ramez 2021], and generic clustering optimizations are rarely tailored for hybrid, multi-stage filtering in resource-constrained scenarios [Stewart and Al-Khassaweneh 2022].

This paper differentiates itself by proposing an automated, task-oriented approach that integrates web scraping, deep learning embedding extraction, and a consensus-driven hybrid outlier removal process. Unlike the reviewed works, this methodology performs multivariate analysis on high-level feature vectors to ensure semantic consistency and visual homogeneity. By combining these stages into a single, low-cost pipeline, the proposed method bridges the gap between raw data acquisition and the generation of high-quality datasets, specifically optimized for lightweight models in domain-specific applications.

3. Methodology

The methodology applied in this work is organized into four main stages, as illustrated in the blocks of Figure 1: (A) image collection, (B) dataset assessment, (C) data cleaning, and (D) classifier training. The first stage, described in Subsection 3.1, involved the automatic collection of images from the internet and the initial structuring of the dataset. In the dataset assessment stage, detailed in Subsection 3.2, duplicate and corrupted images were removed from the dataset, and its quality was measured. The third stage, further detailed in Subsection 3.3, consisted of the automatic cleaning of the dataset, in which outliers found through the union of multiple techniques were removed. Finally, Subsection 3.4 describes the final stage of the classifier training methodology, in which an experimental classification model was optimized and trained on the dataset generated in this case study.

3.1. Image Collection

The construction of the dataset for classifying geometric shapes in real-world images necessitated a systematic collection methodology, encompassing search term definition and structured storage. This process involved four main stages: (i) defining search data, (ii) linguistically structuring queries, (iii) implementing web scraping, and (iv) organizing the

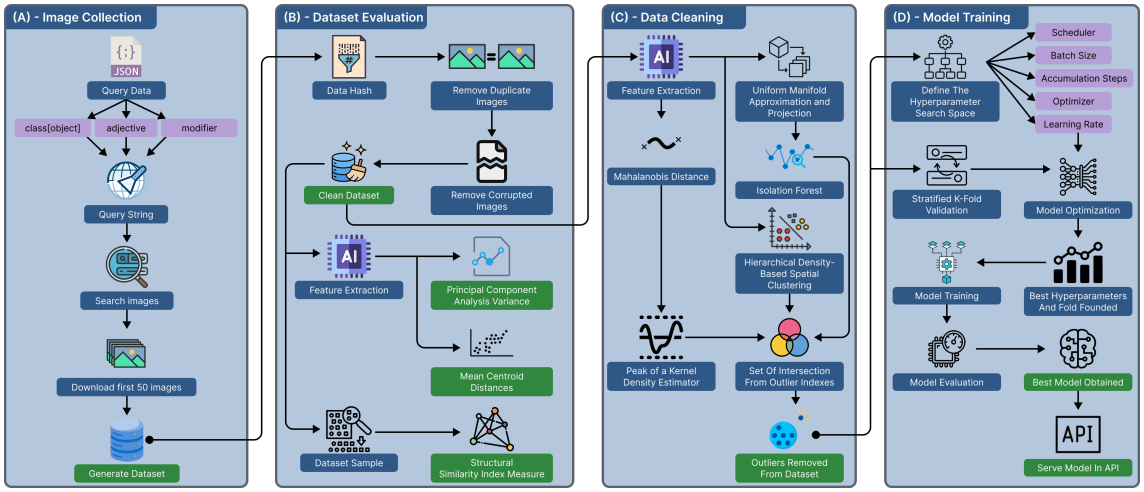

Figure 1. Proposed methodological pipeline for structured image scraping, automated assessment, hybrid outlier removal, and, finally, classifier training and validation.

dataset. The query parameters were meticulously defined across four elements: (i) geometric classes (square, rectangle, triangle, circle, ellipse, parallelogram, trapezoid, pentagon, hexagon); (ii) descriptive adjectives (e.g., color, texture); (iii) contextual modifiers (e.g., scenery, perspective); and (iv) a noun representing the object of the class.

Each query string was formatted for search engine compatibility by replacing spaces with ”+” and composed via controlled pseudo-random combinations, as exemplified in Table 1. To enhance diversity, mechanisms for sampling terms without replacement, random inclusion/exclusion of modifiers, and duplicate removal were implemented, ensuring a broad and varied dataset for robust model training.

Table 1. Examples of combinations of terms used to generate search strings.

Class	Adjective	Noun	Modifier	Search String
triangle	yellow	ruler	on a table	yellow+triangle+ruler+on+a+table
circle	irregular	plate	inside a container	irregular+circle+plate+inside+a+container
pentagon	rusty	road sign	side view	rusty+pentagon+road+sign+side+view

Image collection was performed using headless Chrome and Selenium[2] via DuckDuckGo[3], selected for its scraping flexibility, result diversity, and support for open-license filters to ensure a legally compliant dataset. Post-collection, images were standardized to a 224x224 RGB format using the Pillow library to enable ResNet50 feature extraction, incorporating automated error handling for corrupted files. Finally, the data were systematically organized into class-specific directories (e.g., `dataset/triangle/`) with descriptive filenames combining the source query and index, streamlining their use in supervised learning tasks.

[2]https://www.selenium.dev
[3]https://duckduckgo.com

3.2. Dataset Evaluation

The assessment phase (Figure 1-B) objectively evaluates the raw dataset through a validation pipeline that verifies file integrity. After eliminating redundancies via SHA-256 cryptographic hashing, validated images undergo feature extraction using a ResNet50 CNN pre-trained on ImageNet1K V2 [4]. By capturing 2048-dimensional vectors from the penultimate layer, the method transforms visual content into semantic representations, establishing a robust foundation for analyzing intra-class similarity and dispersion across the nine geometric categories.

To quantify visual diversity, Principal Component Analysis (PCA) is applied to these feature vectors using the scikit-learn library, focusing on the cumulative explained variance of the first 10 principal components. This intra-class approach diagnoses group-specific patterns: concentrated variance (> 0.8) indicates high redundancy, while more uniform distributions ($[0.4 - 0.6]$) suggest greater heterogeneity and richness. Beyond serving as a metric for internal cohesion and representativeness, PCA acts as a dimensionality reduction technique that facilitates the identification of class-specific biases and latent patterns, ensuring a balanced baseline of visual features before subsequent filtering.

Class cohesion is evaluated using Euclidean distances to the centroid (mean feature vector), where smaller values indicate higher internal consistency within the feature space. To complement this, the Structural Similarity Index (SSIM) assesses visual homogeneity by incorporating luminance, contrast, and local structure. Goyal et al. (2015) highlights SSIM's sensitivity to structural distortions; for instance, an image with Gaussian noise yielded a PSNR of 39.67 dB with an SSIM of only 0.136, while compressed images maintained a high SSIM (0.872) despite similar PSNR levels (40.39 dB). To handle high data volumes efficiently, a random sampling approach is used to estimate consistency without exhaustive pairwise comparisons.

Following intra-class analysis, the pipeline consolidates metrics—including PCA explained variance, centroid distances, and SSIM indices—through arithmetic means to generate global descriptive statistics. This quantitative summary provides an objective baseline for evaluating the dataset's overall diversity and structural homogeneity, serving as a critical input for the subsequent outlier removal stage. Furthermore, the automated implementation of this process ensures reproducibility and provides comparable metrics across different dataset versions, facilitating a standardized and reliable development cycle for high-quality visual data.

3.3. Data Cleaning

This phase focuses on removing outliers to preserve dataset integrity. We employed a hybrid consensus approach combining three complementary methods: density-based clustering, multivariate statistical analysis, and isolation forests (Figure 1-C). Inputs consisted of the normalized 2048-dimensional ResNet50 feature vectors described in Subsection 3.2, ensuring representation consistency. Prior to detection, feature magnitudes were unified via L2 regularization, followed by Uniform Manifold Approximation and Projection (UMAP) for non-linear dimensionality reduction. UMAP was selected for its superiority in preserving global structure in high-dimensional manifolds [Rabasovic et al. 2023].

[4]https://www.tensorflow.org/datasets/catalog/imagenet_v2

HDBSCAN was utilized for density-based detection due to its ability to identify variable-density clusters without pre-defined global thresholds. Adaptive parameterization scaled the minimum cluster size relative to the sample count per class, flagging points outside stable regions as noise. Complementarily, Mahalanobis distance identified statistical outliers by accounting for multivariate correlations [Rivas et al. 2020]. Thresholds were automated via Kernel Density Estimation (KDE), setting the cutoff at the distribution peak plus one standard deviation. Finally, Isolation Forest was applied to isolate anomalies via recursive partitioning [Al Farizi et al. 2021]. Operating in the UMAP-reduced space, the algorithm used class-specific adjustments for tree count and contamination factors, defining dynamic thresholds based on anomaly scores for automated filtering.

3.4. Model Training

The training flow (Figure 1-D) was executed on a cluster equipped with two Intel Xeon Gold 6530 CPUs and two NVIDIA A100-SXM4 GPUs. Data preparation involved a stratified $k = 5$ cross-validation split and preprocessing with EfficientNet-B0 normalization. To enhance robustness and invariant feature learning, data augmentation techniques—including random horizontal flips, rotations, perspective distortions, and Gaussian blur—were applied to the training set. This process ensured high data diversity and reduced overfitting, while reproducibility was maintained by storing fixed training and validation indices.

The proposed architecture utilizes a pre-trained EfficientNet-B0 backbone [Tan and Le 2019] coupled with a lightweight edge feature extractor, fused via an Efficient Channel Attention (ECA) module to prioritize salient information [Zhang et al. 2024]. Hyperparameter optimization was conducted via Optuna, exploring a logarithmic learning rate (10^{-6} to 10^{-2}), batch sizes up to 512, and Adam/AdamW optimizers. This automated search, alongside gradient accumulation (1 to 8 steps) and weight decay optimization, aimed to maximize the validation F1-Score while maintaining computational efficiency.

Training prioritized robustness through learning rate scheduling and early stopping based on the F1-Score to prevent overfitting [Wen et al. 2021]. Model performance was validated using weighted F1-Score, accuracy, and precision, complemented by a confusion matrix for per-class analysis. To ensure suitability for resource-constrained environments, the final evaluation included GFLOPs and inference time measurements, providing a comprehensive assessment of both predictive power and operational feasibility on mobile hardware.

4. Case Study

This case study aims to demonstrates the practical applicability of the proposed methodology through the automated classification of two-dimensional geometric shapes. The central motivation stems from the need to optimize the Geometa educational platform, dedicated to teaching geometry via Augmented Reality (AR) and Virtual Reality (VR) in environments with limited infrastructure. This initiative focuses on enhancing pedagogical tools for basic education Mathematics by using AI to recognize patterns within streaming data from local objects and landscapes.

In this context, a lightweight computational model is proposed to identify shapes such as triangles, squares, and circles directly from mobile device cameras. The choice

of an efficient, low-cost model is justified by the technological reality of regions with unstable or nonexistent internet access, where a significant portion of households remains offline and available devices have limited hardware resources [Vieira 2022]. By prioritizing offline operation and low computational demand, the solution ensures that AI-driven educational tools are functional even in areas where cloud-based processing is unfeasible.

To enable this solution, a custom dataset of real-world objects was constructed and refined through the proposed outlier removal stage. The modeling process utilized deep learning techniques adapted for low-capacity environments, specifically designed for direct deployment on smartphones. Ultimately, this case study exemplifies the potential of the method to expand the reach of educational technologies, making them more inclusive and adapted to diverse realities. The integration of the model represents a significant advancement in using AI as a pedagogical tool to promote learning autonomy and democratize access to mathematical knowledge in resource-constrained settings.

5. Results and Discussion

Data collection utilized 6000 search strings based on 191 everyday objects, yielding 106,048 images across nine classes by capturing up to 50 images per query. This raw dataset was refined through a multi-method outlier detection protocol (Table 2), where Mahalanobis identified an average of 4528.14 outliers per class, followed by HDBSCAN (2030.26) and Isolation Forest (1444.82). By taking the union of these results, averaging 5883.85 total outliers per class, the approach ensured the robust identification of atypical samples, significantly enhancing dataset consistency and representativeness for subsequent modeling phases.

Table 2. Quantity of outliers identified per class with the employed methods.

Classes	Mahalanobis	HDBSCAN	Isolation Forest	Total Outliers
circle	4583	1186	1623	6092
ellipse	3888	22	1354	4846
hexagon	2503	611	880	3379
parallelogram	3037	4033	944	5905
pentagon	2982	278	927	3750
rectangle	5006	110	1462	6041
square	4437	95	1304	5256
trapezoid	3280	323	1108	4275
triangle	3447	0	1007	4227
mean (± std)	4528.14 (± 843.36)	2030.26 (± 1290.48)	1444.82 (± 266.04)	5883.85 (± 1444.82)

The removal of outliers demonstrated a quantifiable impact on the quality of the dataset for the proposed case study, as evidenced by the comparative metrics displayed in Figure 2. In the initial state, the dataset presented a mean PCA variance of 0.2586, a mean distance to the centroid of 11.8642 ($\pm$ 1.9691), and a mean SSIM of 0.2971 ($\pm$ 0.1736), indicating significant dispersion and low intra-class structural consistency. After the filtering process, an increase of 25.6% in the mean PCA variance (0.3247) was observed, suggesting greater diversity of valid features. The 5.6% reduction in the mean distance to the centroid to 11.1957 ($\pm$ 1.7291) reflects a more cohesive distribution of

the samples in the latent space, while the 21.6% increase in the mean SSIM (0.3613 ± 0.1891) attests to the greater visual homogeneity of the classes.

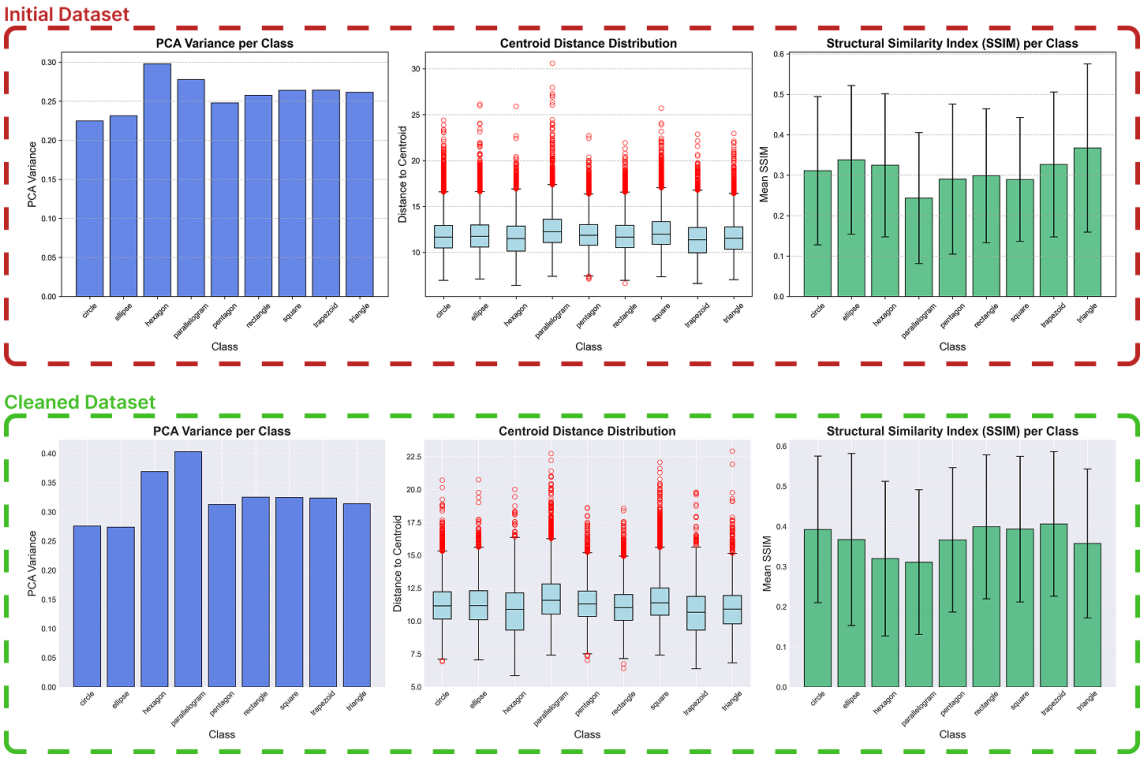

Figure 2. Comparison between the initial generated dataset and after the removal of outliers per class.

The hyperparameter optimization process involved 100 automated searches with 5-fold cross-validation to maximize the validation F1-Score. The optimal configuration comprised a batch size of 256, a learning rate of approximately 0.0032, the AdamW optimizer, StepLR scheduler, 2 accumulation steps, and a weight decay of 6.21×10^{-4}. As illustrated in Figure 3, peak performance was concentrated within the 10^{-3} learning rate range and the 10^{-4} to 10^{-3} weight decay range, effectively balancing regularization with convergence stability. These findings underscore the importance of systematic hyperparameter exploration and stable batch sizes in achieving the high discriminative capacity and robust generalization observed in the final model.

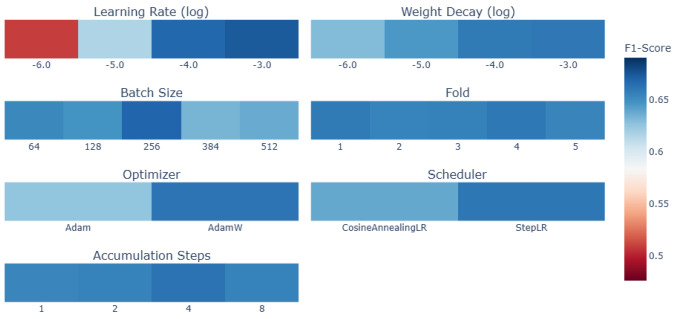

Figure 3. Mean F1-Score obtained for each value range of the selected hyperparameters for the classifier optimization.

During the training process of the developed model, a consistent evolution of performance was observed, as illustrated in Figure 4. The F1-Score metric was used to track the performance on both the training and validation sets, revealing a stable trend of learning and generalization. The difference between the training and validation curves remained controlled throughout the epochs, which indicates that the model did not suffer from significant overfitting. This behavior suggests that the adopted strategies, such as the use of regularization via weight decay, dropout, and gradient accumulation, were effective in ensuring the robustness of the learning process, even with the focus on computational efficiency.

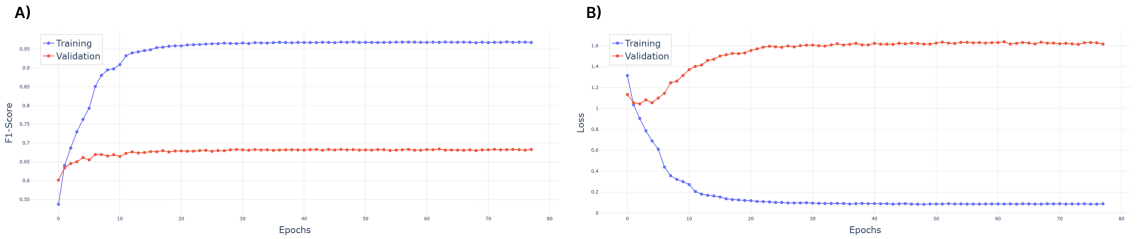

Figure 4. F1-Score obtained on the training and validation sets during the training of the developed experimental model.

The variation of the learning rate during the training epochs was conducted with the StepLR scheduling policy. This policy promotes periodic reductions of the learning rate, allowing the model to make fine adjustments in the last stages of training. Initially, the higher learning rate facilitates rapid convergence, while the subsequent reductions favor stability in the approximation of local minima. The controlled variation of the learning rate was essential to achieve a balance between convergence speed and final model quality [Kim 2024], contributing to the good results observed in the evaluation stage.

During the final evaluation stage, the classification model achieved an expressive performance on the test set, with an F1-Score of 92.5%, an accuracy of 92.5%, and a mean inference time of approximately 0.49 seconds per image, reflecting a rate of 2.02 inferences per second. The computational cost measured in GFLOPs was approximately 0.64, which indicates a processing-efficient model, suitable for applications on devices with limited computational resources. Figure 5 presents the confusion matrix with the correct predictions and errors per class, evidencing a balanced accuracy rate among the different geometric shapes.

Despite the positive results, a significant discrepancy was observed between the performance obtained during validation (mean F1-Score of 68%, on the red curve of Figure 4-A) and the final results on the test set. A possible explanation for this difference lies in the absence of stratification during the data separation process for validation. Without stratification, the distribution of classes among the folds may have been imbalanced, leading the model to be evaluated in scenarios non-representative of the dataset's real distribution. This may have caused an underestimation of the performance during the validation phase, masking the model's real capacity to generalize. The use of stratified validation in future works concerning the pipeline will be implemented, in order to ensure a more faithful assessment of the original class distribution.

Considering the results obtained, the proposed method presented a consistent and

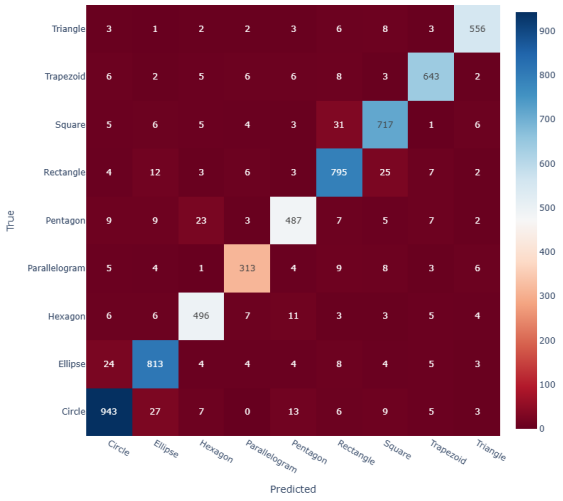

Figure 5. Confusion matrix comparing the true class with the one predicted by the classifier on the test set.

effective performance in the context of the case study. The automated dataset curation stage, combined with the removal of outliers by consensus among three distinct techniques, resulted in a cleaner and more representative dataset, which was directly reflected in the quality of the trained model. The final precision of 92.5% on the test set, accompanied by an F1-score with a similar value, evidenced the assertiveness of the model built based on the refined dataset. Furthermore, the significant reduction in variance between classes and the control over noisy samples contributed to a more robust generalization. Such results reinforce the applicability of the method in real scenarios of low connectivity and infrastructure, especially due to the low GFLOP cost.

6. Conclusion

This work aimed to develop an automated and structured method for the construction, curation, cleaning, and evaluation of visual datasets. The methodology integrates (i) lightweight visual feature extraction; (ii) a consensus-driven outlier removal protocol utilizing Mahalanobis distance, HDBSCAN, and Isolation Forest; (iii) the structuring of datasets for enviroments with limited computational resources; and (iv) a robust hyperparameter optimization flow with stratified cross-validation. This integrated approach ensures the generation of high-quality, task-specific data while minimizing manual intervention and computational overhead.

Empirical results from the the study case demonstrate the method's efficiency, with the final model achieving 92.5% accuracy and F1-score while maintaining a reduced computational footprint of 0.63 GFLOPs and an inference rate of 2 FPS. The pipeline proved capable of producing representative datasets and models functional in offline, resource-constrained scenarios. The discriminative capacity across nine geometric categories validates the feasibility of deploying such AI solutions in basic education contexts where conventional, cloud-reliant infrastructures are unavailable.

Despite these advancements, limitations include the lack of specific public bench-

marks and the need for more sophisticated duplicate detection beyond cryptographic hashes, such as perceptual hashing or feature similarity analysis. Future research will explore expanding this framework to text and tabular data, as well as more complex computer vision tasks like object detection and semantic segmentation. Additionally, integrating active learning and interactive curation interfaces could further refine the reliability of automated datasets and their impact in interdisciplinary or critical domains.

References

Adnan, M. M., Rahim, M. S. M., Rehman, A., Mehmood, Z., Saba, T., and Naqvi, R. A. (2021). Automatic image annotation based on deep learning models: a systematic review and future challenges. *IEEe Access*, 9:50253–50264.

Aghabagherloo, A., Abadi, A., Sarkar, S., Dasu, V. A., and Preneel, B. (2025). Impact of data duplication on deep neural network-based image classifiers: Robust vs. standard models. *arXiv preprint arXiv:2504.00638*.

Al Farizi, W. S., Hidayah, I., and Rizal, M. N. (2021). Isolation forest based anomaly detection: A systematic literature review. In *2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*, pages 118–122. IEEE.

Gong, Y., Liu, G., Xue, Y., Li, R., and Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, 162:107268.

Goyal, M., Lather, Y., and Lather, V. (2015). Analytical relation & comparison of psnr and ssim on babbon image and human eye perception using matlab. *International Journal of Advanced Research in Engineering and Applied Sciences*, 4(5):108–119.

Guo, L. L., Pfohl, S. R., Fries, J., Posada, J., Fleming, S. L., Aftandilian, C., Shah, N., and Sung, L. (2021). Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. *Applied clinical informatics*, 12(04):808–815.

Habuza, T., Navaz, A. N., Hashim, F., Alnajjar, F., Zaki, N., Serhani, M. A., and Statsenko, Y. (2021). Ai applications in robotics, diagnostic image analysis and precision medicine: Current limitations, future trends, guidelines on cad systems for medicine. *Informatics in Medicine Unlocked*, 24:100596.

Kadhim, Y. A., Khan, M. U., and Mishra, A. (2022). Deep learning-based computer-aided diagnosis (cad): Applications for medical image datasets. *Sensors*, 22(22):8999.

Kim, T.-G. (2024). Hyperboliclr: Epoch insensitive learning rate scheduler. *arXiv preprint arXiv:2407.15200*.

Li, X., Li, T., Li, S., Tian, B., Ju, J., Liu, T., and Liu, H. (2023). Learning fusion feature representation for garbage image classification model in human–robot interaction. *Infrared Physics & Technology*, 128:104457.

Nassif, A. B., Talib, M. A., Nasir, Q., and Dakalbab, F. M. (2021). Machine learning for anomaly detection: A systematic review. *Ieee Access*, 9:78658–78700.

Rabasovic, M., Pavlovic, D., and Sevic, D. (2023). Analysis of laser ablation spectral data using dimensionality reduction techniques: Pca, t-sne and umap. *Contrib. Astron. Obs. Skalnaté Pleso*, 53(3):51–57.

Rivas, D. V., Pino, M. M., Pérez, Y. P., Rivas, S. V., Torres, O. G., Fernández, V. C., and Ysa, R. S. (2020). Comparison of distance methods for detection of atypical observations in monthly precipitation series. *Revista de Climatología*, 20.

Rodríguez-Rodríguez, J. A., López-Rubio, E., Ángel-Ruiz, J. A., and Molina-Cabello, M. A. (2024). The impact of noise and brightness on object detection methods. *Sensors*, 24(3):821.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Sharma, P., Ninomiya, T., Omodaka, K., Takahashi, N., Miya, T., Himori, N., Okatani, T., and Nakazawa, T. (2022). A lightweight deep learning model for automatic segmentation and analysis of ophthalmic images. *Scientific reports*, 12(1):8508.

Stewart, G. and Al-Khassaweneh, M. (2022). An implementation of the hdbscan* clustering algorithm. *Applied Sciences*, 12(5):2405.

Tan, J., Hou, B., Day, T., Simpson, J., Rueckert, D., and Kainz, B. (2021). Detecting outliers with poisson image interpolation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 581–591. Springer.

Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946.

Thota, P. and Ramez, E. (2021). Web scraping of covid-19 news stories to create datasets for sentiment and emotion analysis. In *Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference*, pages 306–314.

Vieira, N. (2022). Accessibility in the legal amazon: Digital solutions. *Climate Policy Initiative*.

Wen, L., Gao, L., Li, X., and Zeng, B. (2021). Convolutional neural network with automatic learning rate scheduler for fault classification. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12.

Xiao, K. and Ni, T. (2024). Computer-aided industrial product design based on image enhancement algorithm and convolutional neural network. *Computer-Aided Design and Applications*, 21(1).

Xu, Z., Liu, Z., Yan, Y., Liu, Z., Yu, G., and Xiong, C. (2024). Cleaner pretraining corpus curation with neural web scraping. *arXiv preprint arXiv:2402.14652*.

Zhang, Y., Zhan, Q., and Ma, Z. (2024). Efficientnet-eca: A lightweight network based on efficient channel attention for class-imbalanced welding defects classification. *Advanced Engineering Informatics*, 62:102737.