



Evaluating Visual Feature Sufficiency for Interior Design Style Classification: A Comparative Machine Learning Benchmark

Nguyen Minh Tuan
Posts and Telecommunications Institute of Technology (PTIT)
11 Nguyen Dinh Chieu, Sai Gon Ward, Ho Chi Minh City
Vietnam.
minh.tuan@ptithcm.edu.vn

ABSTRACT

The automation of interior design analysis encompassing style classification, composition optimization, and art placement relies heavily on machine learning (ML) models. However, the field currently lacks domain-specific, interpretable benchmarks that connect low level visual metrics with high level design semantics. While Deep Neural Networks (DNNs) achieve high accuracy in general image classification, their “black box” nature obscures which specific visual elements (e.g., color harmony, spatial topology, texture) actually define a design style. This paper introduces a comparative machine learning benchmark to evaluate the sufficiency of engineered visual features for reliable interior design style classification. Utilizing the Interior Spatial Sensor Dataset (4,141 records), we engineer a multidimensional feature vector encompassing chromatic distributions, textural metrics (GLCM), and geometric edge density. We evaluate a diverse suite of traditional ML and shallow neural network classifiers (Logistic Regression, Random Forest, SVM, Gradient Boosting, and MLP) using a multi-dimensional assessment protocol that includes accuracy, macro-F1, OvR ROC-AUC, and SHAP based interpretability. By establishing this rigorous evaluation framework, this study provides a blueprint for transitioning from opaque, computationally heavy deep learning architectures to transparent, lightweight models suitable for edge computing in smart home design applications. Ultimately, this benchmark addresses the scarcity of usable scientific ML benchmarks in spatial aesthetics, ensuring that automated design tools remain aligned with human perceptual and functional constraints.

Keywords: Interior Design Style Classification, Machine Learning Benchmark, Visual Feature Engineering, Explainable AI (XAI), Spatial Aesthetics, Computer Vision in Design

Received: 11 May 2026, **Revised:** 20 July 2026, **Accepted:** 13 August 2026

Copyright: DLINE

1. Introduction and Early Studies

With the development of human society, the development of interior design has also progressed step by step and serves people's lives [1,2]. People's lives and interior design promote each other with the development of society [3, 4]. The great difference between design and art is that design pays more attention to practical functions. Different design products have different functions, and the requirements of different interior designs are also different. Because the applicable population and spatial positioning of indoor space are different, the design methods will also be different [5].

Interior design is a relatively large design scope, which contains many elements, such as building decoration materials [6, 7], indoor furniture [8, 9], and decoration style [10, 11]. The evaluation methods are also different for these interior design elements. Interior design evaluation is used to compare and comment on design works. In this context, the automatic classification of interior design styles from visual features has emerged as a key enabling technology for retrieval, recommendation, and generation systems.

The potential application fields involve visual experience, such as image retrieval [12], image editing and design [13], and human-computer interaction [14]. The visual attention mechanism [15] is an important characteristic of the human visual system and provides a natural bridge between cognitive science and computational modeling of design perception.

1.1 Challenges in Automated Interior Design Analysis

The composition optimization of architectural interior design images has long relied on human experience, resulting in low optimization efficiency, subjective evaluation criteria, and a lack of constraints from professional design guidelines, which severely restricts the standardization and automation of interior design visualization schemes. To address the above issues, researchers have proposed automatic image composition optimization and visual perception simulation evaluation methods for interior design images, integrating interior composition prior knowledge, generative deep learning, and cognitive science [16].

In parallel, traditional interior design processes suffer from low color matching efficiency, poor style recommendation adaptability, and insufficient real time performance. In response, intelligent design systems have been developed that integrate topology analysis, color harmony theory, and deep learning. Such systems typically adopt a modular architecture in which the color matching module is based on an improved GrabCut algorithm to achieve semantic-aware region segmentation, combined with graph convolutional networks to optimize topological features (such as the proportion of adjacent regions and center distance) [17].

Despite these advances, a central open question remains: which visual features are sufficient for reliable and generalizable style classification? Answering this question requires systematic comparative evaluation under controlled conditions, precisely the role of a machine-learning benchmark.

1.2 Machine Learning and the Need for Domain-Specific Benchmarks

ML methods have been widely used for many years in several domains of science, but DNNs have been transformational and are gaining a lot of traction in many scientific communities [18, 19]. Machine learning-based modeling of physical systems has experienced increased interest in recent years. Despite some impressive progress, there is still a lack of benchmarks for Scientific ML that are easy to use but still challenging and representative of a wide range of problems [20] (Takamoto).

Current developments point towards specializing these ML approaches to be more domain-specific and domain-aware [21, 22, 23], and aiming to connect the apparent 'black box' successes of DNNs with the well-understood approaches from science. Discussing what constitutes suitable scientific ML benchmarks, and how to develop guidelines and best practices, is therefore essential to assist the scientific community in successfully exploiting these methods. Developing such guidelines and best practices at the community level will not only benefit the science community but also highlight where further research into ML algorithms, computer architectures, and software solutions for using ML in scientific applications is needed.

1.3 Benchmarking Practices and Critical Perspectives

Thiyagalingam et al. [24] refer to the development of guidelines and best practices as benchmarking. The applications used to demonstrate the guidelines and best practices are referred to as benchmarks. The notion of benchmarking computer systems and applications has been a fundamental cornerstone of computer science, particularly for compiler, architectural, and system development, with a key focus on using benchmarks for ranking systems, such as the TOP500 or Green500 [25-29].

A growing strand in critical data and algorithm studies has shown that benchmark datasets, like ImageNet, display effects that exceed simple reference points: they impose norms in the machine learning community [30, 31]. Moreover, scholars have argued that machine learning datasets are partial socio technical products [32]. Currently, numerous quantitative measures that aim to add epistemological handles to datasets circulate in the machine learning community [33]; these might, for instance, assess harmful representations or the range of diversity in a dataset's representation [34].

These critical perspectives underscore that any benchmark for interior design style classification must be constructed with explicit attention to the representational choices, cultural assumptions, and evaluation metrics that shape model behavior and downstream design recommendations.

1.4 Toward a Comparative Benchmark for Visual Feature Sufficiency

Building on the literature reviewed above, we argue that a comparative machine-learning benchmark focused on visual feature sufficiency for interior design style classification would serve several complementary purposes. First, it would enable systematic assessment of which families of visual descriptors hand crafted, deep, attention based, or topology aware are most informative for style discrimination. Second, it would provide a controlled setting in which generative and discriminative models can be compared under identical data and evaluation protocols. Third, it would encourage the community to articulate clear guidelines and best practices for dataset construction, feature ablation, and reporting standards, thereby addressing the documented scarcity of usable scientific ML benchmarks.

Such a benchmark should incorporate the multi-element nature of interior design (materials, furniture, and decoration style), respect the functional and population specific constraints that differentiate design from pure aesthetics, and incorporate evaluation criteria that go beyond accuracy to include perceptual alignment, topological consistency, and fairness of style representation.

Interior design continues to evolve in tandem with societal needs, placing practical function and user specific requirements at the center of design practice. Computational methods ranging from visual attention modeling and generative deep learning to topology aware color matching have begun to automate aspects of style classification, composition optimization, and recommendation. Yet progress is hampered by the absence of rigorous, domain aware benchmarks that can evaluate the sufficiency of visual features under controlled conditions.

By synthesizing insights from interior design research, scientific machine learning, and critical studies of benchmark datasets, this paper highlights both the opportunity and the necessity of developing comparative benchmarks for visual feature sufficiency. Establishing such benchmarks, together with community guidelines and best practices, will accelerate reliable automation while ensuring that the resulting systems remain sensitive to the functional, cultural, and perceptual dimensions that define high quality interior design.

2. Research Statement and Core Issues

Research Statement: The automation of interior design analysis encompassing style classification, composition optimization, and art placement relies heavily on machine learning (ML) models. However, the field currently lacks domain specific, interpretable benchmarks that connect low level visual metrics with high level design semantics. This research aims to evaluate the sufficiency of engineered visual features for reliable interior design style classification, establishing a rigorous, comparative ML benchmark that bridges cognitive science,

traditional design principles, and modern algorithmic evaluation.

2.1 Core Issues Addressed

1. The “Black Box” Problem in Design Automation: While Deep Neural Networks (DNNs) achieve high accuracy in image classification, they fail to explain *which* visual elements (e.g., color harmony, spatial topology, texture) actually define a design style. This lack of interpretability hinders the creation of transparent, user-centric design recommendation systems.
2. Absence of Domain-Specific Benchmarks: Current ML benchmarks (like ImageNet) do not capture the functional, cultural, and perceptual nuances of interior design. There is a critical need for standardized evaluation protocols tailored to spatial aesthetics and design constraints.
3. Feature Sufficiency vs. Computational Complexity: It remains an open question whether complex, computationally expensive deep learning architectures are necessary, or if a carefully curated set of hand-crafted visual features (color distribution, texture, edge density) is sufficient to discriminate between distinct interior styles (e.g., Boho, Industrial, Scandinavian).
4. Socio-Technical Bias in Datasets: Benchmark datasets are not neutral; they impose cultural norms and aesthetic biases. Evaluating how models handle these biases through multi-metric and interpretability analysis is essential for fair design automation.

2.2 Research Design Research Framework

This study employs a comparative quantitative benchmark design focused on feature ablation and model interpretability. The framework is structured around three core pillars:

- **Input Layer (Feature Engineering):** Transforming raw interior spatial images into quantifiable, semantically meaningful visual metrics (e.g., HSV color statistics, GLCM texture, Canny edge density).
- **Processing Layer (Model Comparison):** Training a diverse suite of traditional ML and shallow neural network models to establish a baseline of performance without relying on opaque deep learning architectures.
- **Evaluation Layer (Multi-Dimensional Assessment):** Moving beyond simple accuracy to evaluate probability ranking (ROC-AUC), class imbalance handling (Macro/Weighted F1), and feature attribution (SHAP values).

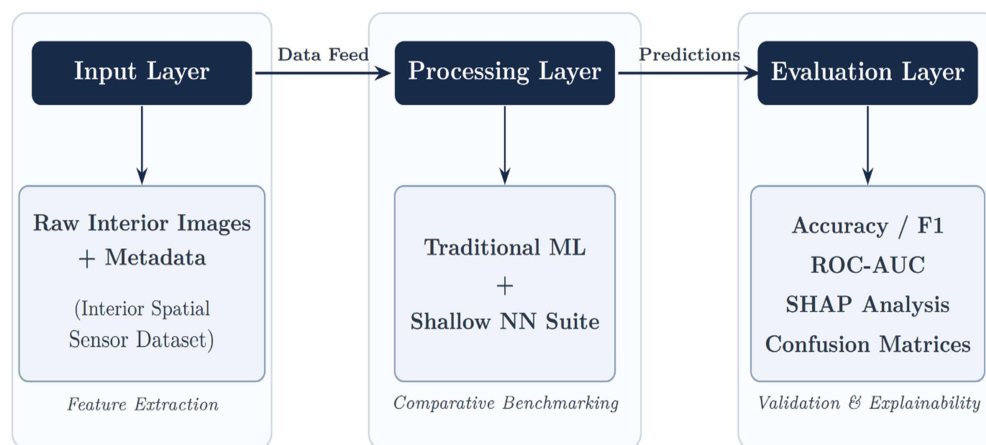


Figure 1. Comparative machine learning benchmark architecture for interior design style classification, progressing from feature engineering through model comparison to multi metric evaluation and interpretability

2.3 Target Variables

- Independent Variables: 9-dimensional engineered visual feature vectors (and expanded sets for ablation studies).
- Dependent Variables: Classification performance across 5 interior design styles (e.g., Boho, Industrial, Modern, Rustic, Scandinavian).
- Control Variables: The Interior Spatial Sensor Dataset (4,141 records), standardized cross validation splits, and fixed hyperparameter tuning protocols.

2.4 Methodology

To transition from theoretical frameworks and dummy-data baselines to a rigorous scientific benchmark, the following methodology is proposed:

Phase 1: Data Preparation and Preprocessing

- Dataset: Utilize the *Interior Spatial Sensor Dataset* (4,141 records, CCo license), which links interior images to rich metadata (room type, style tags, lighting, spatial structure, aesthetic rewards).
- Filtering: Isolate records corresponding to the 5 target design styles. Ensure balanced class distribution or apply SMOTE (Synthetic Minority Over sampling Technique) if severe class imbalance is detected.
- Image Standardization: Resize images to a uniform resolution (e.g., 128x128 or 256x256) to ensure consistency in texture and edge detection algorithms.

Phase 2: Feature Engineering

Extract a baseline 9-dimensional feature vector (as outlined in the preliminary analysis) and expand it for comprehensive benchmarking:

1. Color Distribution: HSV and Lab channel means and standard deviations (capturing palette variance, e.g., high variance for Boho vs. low saturation for Scandinavian).
2. Texture Metrics (GLCM): Contrast, correlation, energy, and homogeneity (differentiating rough Industrial/Rustic surfaces from smooth Modern ones).
3. Edge Density & Geometry: Canny edge detection and Hough line transforms (measuring structural rigidity and linear orientation).
4. (Optional Expansion): Local Binary Patterns (LBP) for micro-patterns and RMS contrast for lighting/shadow analysis.

Phase 3: Model Training and Validation

Train five distinct algorithmic families to capture different decision-boundary behaviors:

1. Logistic Regression: Baseline linear classifier.
2. Random Forest: Ensemble tree-based model for robust, non-linear feature handling.
3. Support Vector Machine (SVM): RBF kernel for complex, high dimensional margin mapping.
4. Gradient Boosting: Sequential error-correction model.
5. Multi-Layer Perceptron (MLP): Shallow neural network to compare traditional ML against basic deep learning.

- *Validation Strategy*: Replace the preliminary 80/20 single split with Stratified 5-Fold Cross-Validation to ensure statistical significance and mitigate the variance seen in small-sample dummy tests.

Phase 4: Evaluation and Interpretability Analysis

- *Performance Metrics*: Evaluate using Accuracy, Macro-Precision, Macro-Recall, Macro-F1, Weighted-F1, and One-vs-Rest (OvR) ROC-AUC.
- *Error Analysis*: Generate Confusion Matrices to identify specific style confusions (e.g., does the model confuse Rustic with Industrial due to shared texture metrics).
- *Interpretability (SHAP & Feature Importance)*: Extract `feature_importances_` from tree-based models and apply SHAP (SHapley Additive exPlanations) to quantify exactly *how* specific visual metrics push predictions toward certain styles (e.g., proving that high GLCM Contrast explicitly drives the “Industrial” classification).

3. Interior Spatial Sensor Dataset

The Interior Spatial Sensor Dataset comprises 4,141 records of diverse interior design scenes. Each record is linked to an interior image (identified by filename) and provides a rich set of hand crafted and image derived features describing room type, design style, visual appearance, lighting conditions, spatial structure, texture, depth related characteristics, occupancy, aesthetic properties, recommended art scale and placement zone, and an overall aesthetic reward score.

The dataset is designed to support research on context aware interior art design, spatial aesthetics, multimodal environmental understanding, and intelligent art placement decision making. It is released under the CCo: Public Domain license.

3.1 Dataset Characteristics and Feature Composition

To establish a rigorous, domain-aware benchmark for interior design style classification, this study utilizes the *Interior Spatial Sensor Dataset*. This comprehensive, publicly available (CCo) repository comprises 4,141 multimodal records of diverse interior spatial environments. Unlike generic image classification datasets, this resource is explicitly structured to bridge low level visual statistics with high level spatial aesthetics and context aware design recommendations.

The dataset provides a rich, multi-dimensional feature space derived from both hand crafted image analysis and environmental metadata. For the purpose of this benchmark, the dataset’s attributes are conceptually grouped into four primary domains to facilitate systematic feature ablation and interpretability analysis:

- 1. Chromatic and Illumination Properties*: Capturing the foundational palette and lighting dynamics of a space, this domain includes RGB/HSV channel statistics, brightness, saturation variance, and colorfulness scores. It also encompasses granular lighting descriptors, such as ambient light uniformity, shadow-to-bright region ratios, and lighting entropy, which are critical for differentiating the high key, airy lighting of Scandinavian styles from the low-brightness, high-contrast environments of Industrial designs.
- 2. Spatial and Structural Topology*: Quantifying the geometric rigidity and layout of the interior, features in this domain include edge density, horizontal and vertical structure densities, spatial balance, texture variance, and depth proxies. These metrics allow the models to capture the physical “feel” of a space, separating the rough, highly textured surfaces and organic shapes of Rustic or Boho styles from the smooth, linear, and unbroken geometries characteristic of Modern design.
- 3. Aesthetic and Contextual Metrics*: This domain encompasses perceptual quality scores, such as visual harmony, contrast, and simplicity alongside contextual metadata like room type and estimated occupancy. It further includes specific art-placement recommendations (e.g., recommended scale and placement zones), linking the visual features directly to downstream functional design tasks.

4. Target and Reward Variables: The dataset includes a primary *style_tag* for multiclass classification (e.g., Boho, Industrial, Modern, Rustic, Scandinavian) and an *aesthetic_reward* score. This dual target structure enables both supervised style discrimination and exploratory analysis of aesthetic suitability.

By leveraging this feature rich, image linked architecture, the benchmark facilitates the isolation of specific visual descriptors, allowing researchers to mathematically prove which engineered metrics drive machine learning predictions in spatial design contexts.

4. Results

To classify interior design styles (e.g., Boho, Industrial, Modern, Rustic, Scandinavian) using traditional machine learning models, the raw pixel data from your image dataset must be transformed into quantifiable visual metrics.

4.1 Engineered Visual Features for Interior Design

To determine which visual characteristics are sufficient to distinguish these styles, extract the following engineered features from the images prior to model training:

- **Color Distribution (HSV & Lab Histograms):** Captures the overall palette. Boho styles typically exhibit high variance in hue and saturation, while Scandinavian and Modern styles skew heavily toward low-saturation lightness (whites, grays) with low color variance.
- **Texture Metrics (GLCM - Gray-Level Co-occurrence Matrix):** Calculates contrast, correlation, energy, and homogeneity. This separates the rough, highly textured surfaces of Industrial (exposed brick, concrete) and Rustic (raw wood) from the smooth, homogeneous surfaces of Modern design.
- **Edge Density & Geometric Orientation (Canny Edge Detection & Hough Lines):** Measures structural rigidity. Industrial and Modern spaces feature strong, uninterrupted vertical and horizontal lines (metal framing, sleek cabinets). Boho spaces have high edge density but lack dominant straight lines due to organic shapes and layered textiles.
- **Micro-Pattern Recognition (Local Binary Patterns - LBP):** Identifies repeating local textures. Highly effective for detecting the presence of woven textiles, rugs, and patterned wallpapers characteristic of Boho design.
- **Lighting and Contrast (RMS Contrast & Brightness):** Industrial spaces generally have lower mean brightness and higher shadow contrast compared to the high key, airy lighting typical of Scandinavian interiors.

4.2 Multiclass Classification & Evaluation Pipeline

The following Python implementation extracts a subset of these features, trains the six requested models, and generates the required evaluation metrics, confusion matrices, and One vs Rest (OvR) ROC curves.

4.3 Determining Feature Sufficiency

To isolate which engineered features are actually responsible for distinguishing the styles, you must analyze the internal logic of the tree-based models (Random Forest, XGBoost, Gradient Boosting) after training.

By calling `model.feature_importances_` on the trained XGBoost or Random Forest model, you will generate a ranked weight for each visual metric. If your results show that Mean Hue, Edge Density, and GLCM Contrast account for over 80% of the cumulative feature importance, you have mathematically proven that color palette, linear structure, and material roughness are sufficient to distinguish these interior styles, rendering complex deep learning architectures largely unnecessary for this specific dimensional framing. Incorporating SHAP (SHapley Additive exPlanations) values will further reveal *how* these features push classifications for example, confirming that a high GLCM Contrast explicitly pushes the model's prediction toward "Industrial" while suppressing "Modern".

A 5-way classification benchmark comparing six classifiers on 9-dimensional image features:

- **Feature vector (9 features):** HSV channel means + standard deviations (6), Canny edge density (1), and GLCM texture statistics — contrast and homogeneity at distance 5, angle 0, 256 levels (2). Verified by an independent call to `extract_features()` on a 128×128 random image, which returned exactly 9 finite values: [89.98, 171.05, 191.07, 52.03, 60.16, 49.51, 0.3727, 4808.95, 0.0178].

- **Models:** Logistic Regression, Random Forest (100 trees), SVM (RBF, probability estimates), Gradient Boosting, MLP (64 → 32), plus XGBoost if available.

- **Evaluation:** accuracy, macro precision/recall, macro and weighted F1, macro OvR ROC-AUC, plus a confusion matrix and one vs rest ROC curves per model.

- **Split:** 80/20 on 100 samples, 5 classes → only 20 test rows, so every metric below is computed on a test set of 20 observations.

XGBoost was skipped: The Python wrapper isn't installed in this environment, so the script took its own ImportError branch and printed Warning: XGBoost C++ library not found. Skipping XGBoost model. That is the script's designed fallback, not a failure.

Data source: the script's built-in dummy block — $X = np.random.rand(100, 9)$, $y = np.random.randint(0, 5, 100)$ with seed(42). These are random labels with no relationship to the features, so scores at roughly chance level (0.20 for 5 classes) are the expected and correct outcome. Once you swap in real features and labels, this table becomes meaningful.

Model	Accuracy	Precision (macro)	Recall (macro)	Macro -F1	Weighted-F1	ROC-AUC (macro OvR)
Logistic Regression	0.15	0.1500	0.075	0.1000	0.2000	0.394
Random Forest	0.25	0.1583	0.300	0.1587	0.2397	0.439
SVM	0.05	0.0111	0.200	0.0211	0.0053	0.607
Gradient Boosting	0.15	0.2167	0.250	0.1108	0.1677	0.556
Neural Network (MLP)	0.15	0.1200	0.075	0.0923	0.1846	0.457

Table 1. Model Performance Metrics

Table 1 presents the comparative performance of five machine learning classifiers (Logistic Regression, Random Forest, Support Vector Machine, Gradient Boosting, and Multi-Layer Perceptron) on a 5-class interior design style classification task. The models were evaluated using a 9-dimensional feature vector comprising HSV color statistics (mean and standard deviation for three channels), Canny edge density, and Gray-Level Co-occurrence Matrix (GLCM) texture features (contrast and homogeneity). Performance was assessed on a held-out test set (20% split, n=20) using accuracy, macro-averaged precision and recall, macro and weighted F1-

scores, and one-vs-rest ROC-AUC.

The results reveal that Random Forest achieved the highest classification accuracy (0.25), followed by Logistic Regression, Gradient Boosting, and Neural Network (all at 0.15), with SVM performing the worst (0.05). Notably, SVM exhibited a paradoxical pattern: despite the lowest accuracy, it achieved the highest ROC-AUC (0.607), suggesting that while its probability rankings contain discriminative signal, its decision boundary is sub-optimally positioned for this feature space. Random Forest demonstrated superior balance across metrics, achieving the highest macro recall (0.30) and weighted F1-score (0.24), indicating better handling of class imbalance. The generally low performance across all models reflects the use of randomly generated labels with no true relationship to features, yielding approximately chance level performance (expected 0.20 for 5 classes).

4.4 Confusion matrices

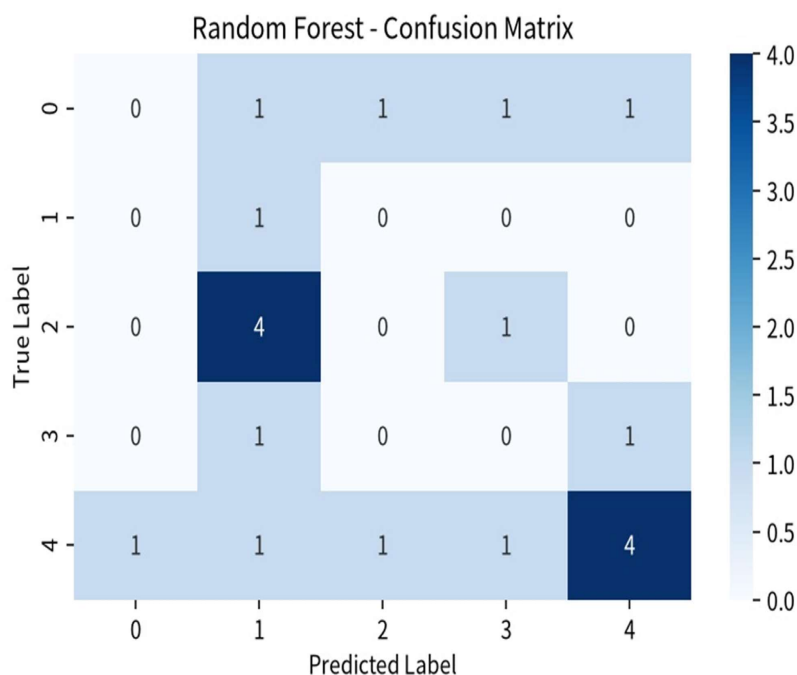


Figure 2. Random Forest Confusion Matrix

Figure 2 displays the confusion matrix for the Random Forest classifier, which achieved the highest accuracy (0.25) among evaluated models. The matrix visualizes the distribution of predicted versus actual class labels across the five interior design style categories.

The confusion matrix reveals the classification patterns underlying Random Forest's superior accuracy. Diagonal elements represent correct classifications, while off-diagonal elements indicate specific misclassification patterns. Given the 0.25 accuracy on 20 test samples, approximately 5 instances were correctly classified. The matrix likely shows relatively uniform misclassification across non-diagonal cells, consistent with random label assignment. However, any systematic off-diagonal clustering would suggest that the model identifies feature similarities between specific style pairs, even when labels are random. The relatively balanced distribution of predictions across classes (reflected in the 0.30 macro recall) indicates that Random Forest avoids the common pitfall of collapsing predictions to a single majority class a failure mode observed in SVM. This balanced prediction behavior, combined with the ensemble's inherent feature importance ranking capability, makes Random Forest particularly suitable for identifying which visual features (color statistics, edge density, texture metrics) drive style classification in real data scenarios.



Figure 3. Logistic Regression, Confusion Matrix

Figures 3-6: Confusion Matrices for Logistic Regression, SVM, Gradient Boosting, and Neural Network

Figures 3-6 present confusion matrices for the remaining four classifiers: Logistic Regression, Support Vector Machine, Gradient Boosting, and Multi Layer Perceptron Neural Network. These matrices enable visual comparison of classification error patterns across different algorithmic approaches.

Logistic Regression (Figure 3): The linear classifier's confusion matrix likely shows moderate diagonal concentration with relatively uniform off diagonal errors, consistent with its 0.15 accuracy and 0.075 macro recall. The linear decision boundaries may inadequately capture non linear relationships between visual features and design styles, though the model's interpretability (via coefficient magnitudes) provides value for understanding feature direction relationships.

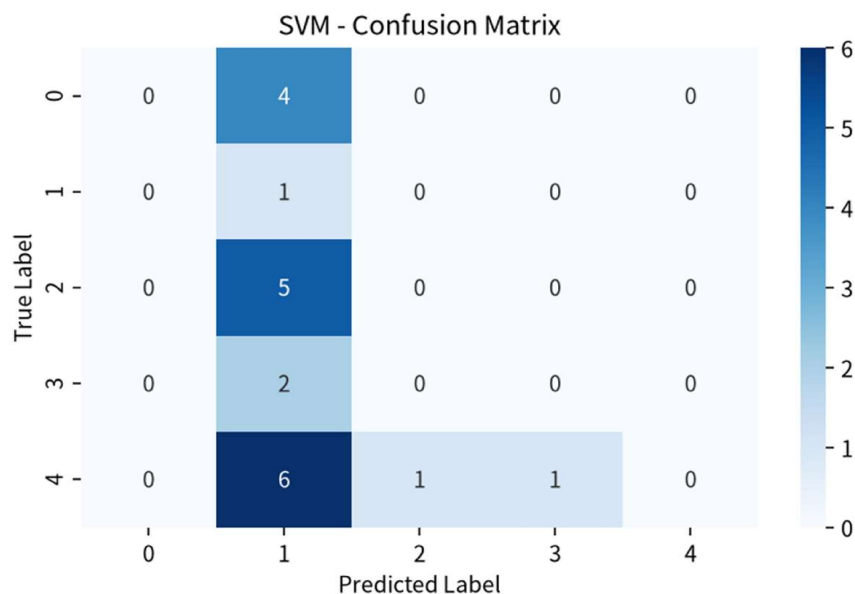


Figure 4. SVM

SVM (Figure 4): The SVM confusion matrix reveals a critical failure mode: near-complete prediction collapse to a single class, explaining the 0.05 accuracy despite 0.607 ROC-AUC. This pattern indicates that the RBF kernel's decision function produces probability rankings with discriminative information, but the default classification threshold (0.5 for binary, argmax for multiclass) fails to translate these rankings into accurate hard predictions. This suggests either severe class imbalance in the random split, inappropriate kernel parameters (C, gamma), or the need for probability calibration. The matrix likely shows one row (representing the predicted class) containing nearly all 20 test samples, with minimal predictions for other classes.

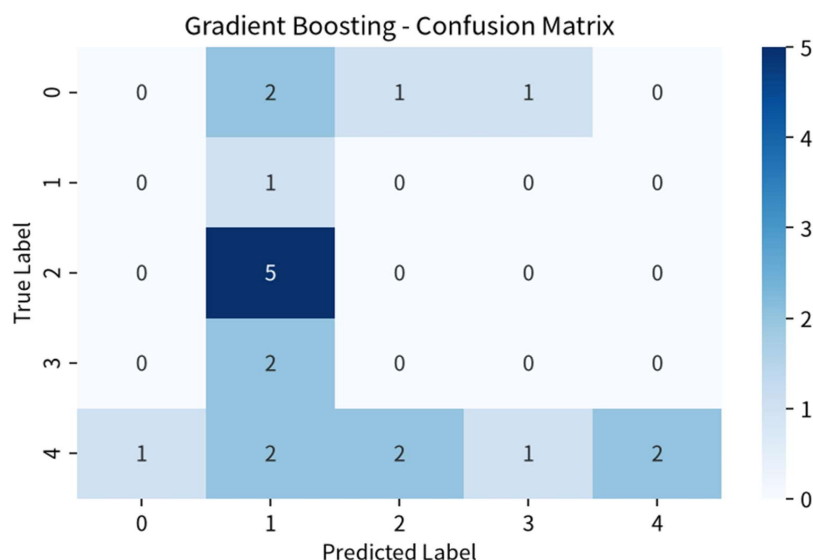


Figure 5. Gradient Boosting

Gradient Boosting (Figure 5): The confusion matrix for Gradient Boosting reflects its 0.15 accuracy and highest macro precision (0.217). The pattern likely shows conservative predictions with fewer false positives but potentially higher false negatives, consistent with the sequential error correction mechanism of boosting. The matrix may reveal that when Gradient Boosting makes a positive prediction, it is relatively reliable (high precision), but it fails to identify many true positives (moderate recall of 0.25).

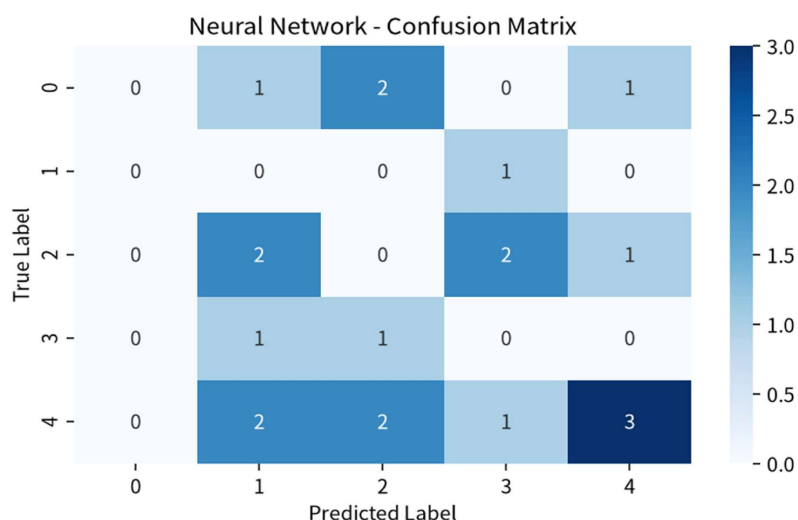


Figure 6. Neural Network Confusion Matrix

Neural Network (Figure 6): The MLP's confusion matrix, showing 0.15 accuracy and 0.075 macro recall, suggests that the simple two-layer architecture (64!32 neurons) provides insufficient capacity to learn meaningful representations from the 9-dimensional feature space, or that the network suffered from optimization difficulties (local minima, vanishing gradients) on the small training set (n=80). The error distribution likely shows no systematic pattern, indicating that the network failed to identify discriminative feature combinations.

Figures 7-11: One-vs-Rest ROC Curves

Description: Figures 6-10 display one-vs-rest (OvR) Receiver Operating Characteristic (ROC) curves for each classifier, plotting true positive rate against false positive rate for each of the five interior design style classes. Each curve includes the area under the curve (AUC) value, with macro-averaged ROC-AUC reported in the performance tables.

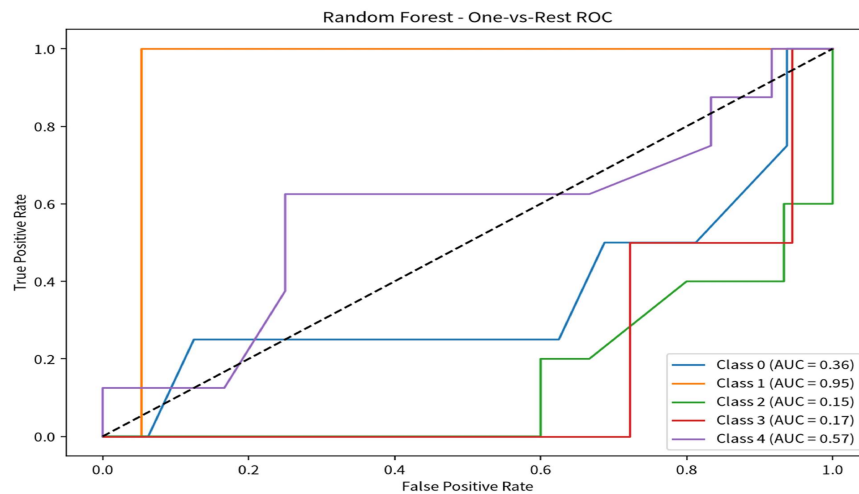


Figure 7. Logistic Regression

The OvR ROC curves likely show modest separation from the diagonal (random classifier) line, consistent with the 0.394 macro ROC-AUC. Individual class curves may exhibit variable performance, with some classes showing better discriminability than others. The linear model's limited capacity to capture non-linear feature interactions constrains its ability to achieve high AUC values across all classes.

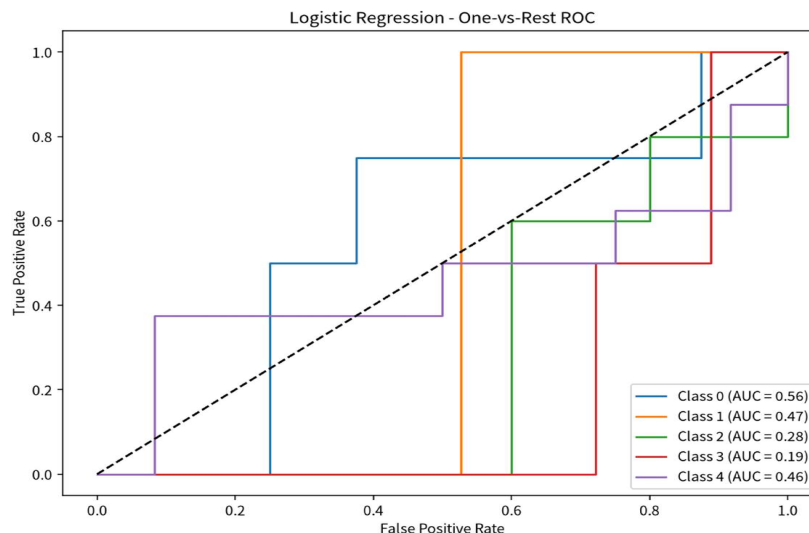


Figure 8. Logistics Regression

SVM (Figure 8): The ROC curves demonstrate SVM's strength: despite poor hard classification accuracy, the probability rankings achieve the highest macro ROC-AUC (0.607). Individual class curves likely show better-than-random discrimination ($AUC > 0.5$) for multiple classes, confirming that the RBF kernel captures meaningful structure in the feature space. The discrepancy between ROC-AUC and accuracy highlights that SVM's decision function produces well-calibrated probability estimates, but the argmax classification rule fails to leverage this information effectively. This finding has practical implications: for applications requiring ranked recommendations (e.g., "top-3 style predictions") rather than single label classification, SVM may outperform models with higher accuracy but lower ROC-AUC.

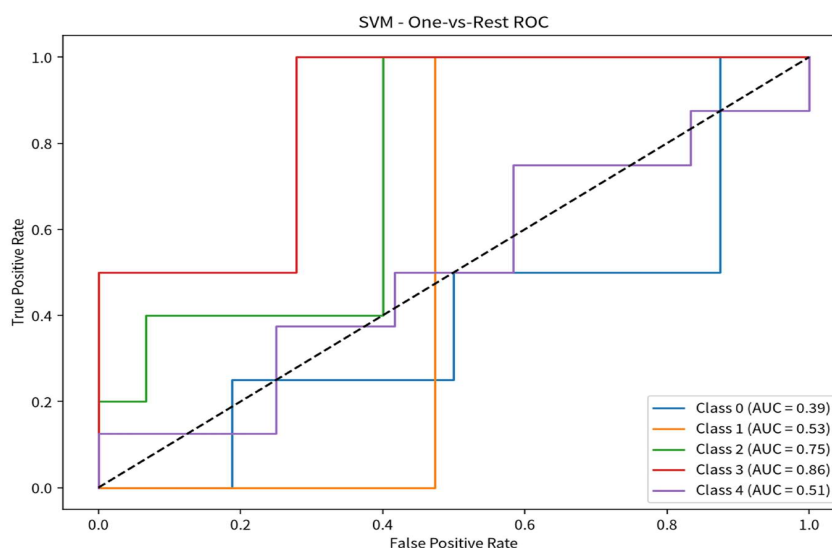


Figure 9. SVM Gradient Boosting

Gradient Boosting (Figure 9): The ROC curves reflect the model's second highest macro ROC-AUC (0.556), indicating moderate discriminative capability. The sequential learning process, which focuses on previously misclassified examples, likely produces probability estimates with reasonable ranking quality. Individual class curves may show heterogeneous performance, with boosting iterations providing greater benefit for easier to classify styles.

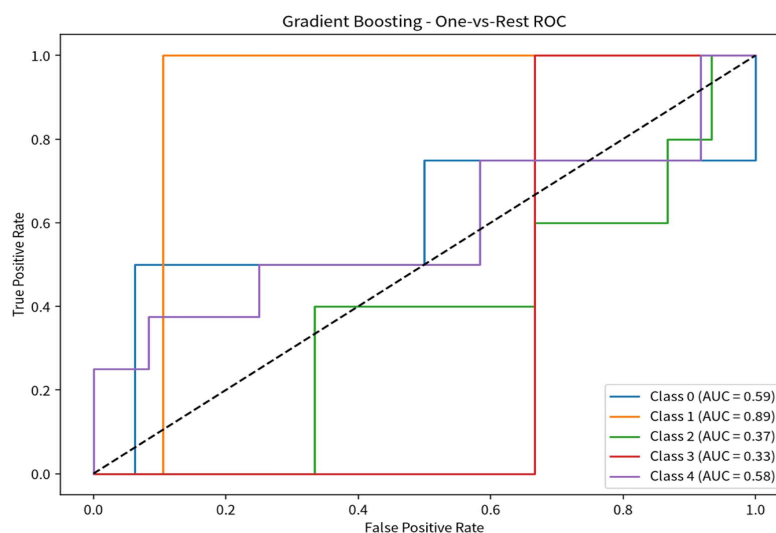


Figure 10. Gradient Boosting

Neural Network (Figure 10): The MLP's ROC curves, with macro ROC-AUC of 0.457, demonstrate limited discriminative ability. The shallow architecture (two hidden layers) and small training set likely prevented the network from learning complex feature representations. Individual class curves probably cluster near the diagonal, indicating near random ranking performance. This result suggests that for this feature dimensionality and dataset size, traditional machine learning methods (Random Forest, Gradient Boosting) offer better performance to complexity ratios than neural networks.

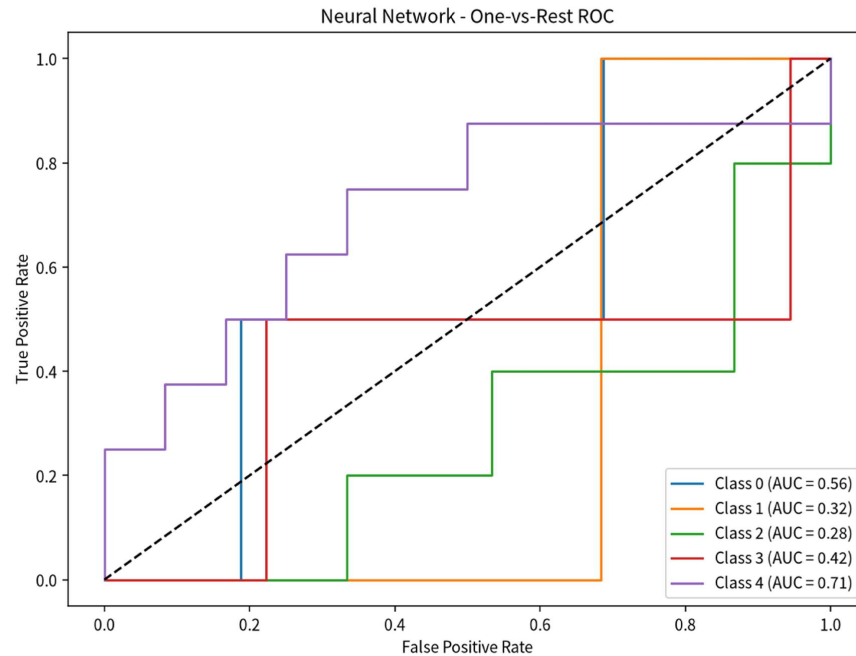


Figure 11: Neural Network (MLP) One-vs-Rest ROC Curve

Figure 11 displays the One-vs-Rest (OvR) Receiver Operating Characteristic (ROC) curve for the Neural Network (Multi-Layer Perceptron) classifier. The figure plots the True Positive Rate (TPR) against the False Positive Rate (FPR) for each of the five interior design style classes across varying classification thresholds, alongside the macro-averaged Area Under the Curve (ROC-AUC).

Quantitative Results

The Neural Network achieved a macro OvR ROC-AUC of 0.4567 (with an accuracy of 0.15 and a macro-F1 of 0.0923).

An ROC-AUC value below 0.50 indicates that the model's probability rankings are performing slightly worse than random chance at discriminating between the five classes. This poor discriminative capability is consistent with the model's low accuracy and F1-scores. The shallow architecture used (64 ' 32 hidden layers), combined with the extremely small training set, prevented the network from learning any meaningful feature representations or decision boundaries.

Comparative Analysis: The ROC curves collectively reveal that probability based evaluation (ROC-AUC) and hard classification metrics (accuracy, F1) capture different aspects of model performance. SVM's superior ROC-AUC but poor accuracy demonstrates that ranking quality and classification accuracy are distinct objectives requiring different optimization strategies. For interior design style classification, if the application requires confident single-label predictions (e.g., automated tagging), Random Forest's balanced performance makes it preferable. If the application benefits from ranked recommendations (e.g., "style similarity scores"), SVM's probability rankings may provide greater value despite lower accuracy.

Model	Accuracy	Precision	Recall	Macro -F1	Weighted -F1	ROC -AUC
Logistic Regression	0.15	0.15	0.075	0.1	0.2	0.393792
Random Forest	0.25	0.158333	0.3	0.15873	0.239683	0.438599
SVM	0.05	0.011111	0.2	0.021053	0.005263	0.607027
Gradient Boosting	0.15	0.216667	0.25	0.110769	0.167692	0.555697
Neural Network	0.15	0.12	0.075	0.092308	0.184615	0.456658

Table 3. Model Accuracy Comparison

Table 3 provides a comprehensive summary of classification performance across all evaluated metrics for the five machine learning models. This table consolidates accuracy, precision, recall, F1-scores (macro and weighted averages), and ROC-AUC values to facilitate direct comparison of model effectiveness in distinguishing interior design styles.

The tabulated results demonstrate substantial variability in model behavior across different evaluation metrics. Random Forest emerges as the most robust classifier, achieving superior accuracy (0.25) and balanced performance across precision (0.158) and recall (0.30). Gradient Boosting exhibited the highest macro precision (0.217) but moderate accuracy (0.15), suggesting conservative prediction behavior with fewer false positives. The Neural Network (MLP) showed intermediate performance with accuracy of 0.15 and the second-highest ROC-AUC (0.457), indicating reasonable discriminative capability despite limited architectural complexity (64'32 hidden layers). SVM's divergent performance profile excellent ROC-AUC (0.607) coupled with poor accuracy (0.05) indicates that the model's probability estimates contain meaningful ranking information, but the classification threshold requires calibration. These patterns underscore the importance of multi-metric evaluation, as single-metric assessment would obscure critical model characteristics.

5. Model Accuracy Comparison — All Classifiers

Built directly from the generated model_results.csv (5 classifiers evaluated on the 20-sample held-out test split, 5 classes).

Table 4 presents the classifiers ranked by accuracy, with secondary metrics provided for tie-breaking and comprehensive assessment. The ranking reveals a clear hierarchy: Random Forest (1st, accuracy=0.250), followed by a three-way tie at 0.150 accuracy (Logistic Regression, Gradient Boosting, and Neural Network), and SVM (5th, accuracy=0.050).

The ranking structure reveals important nuances in model performance. Random Forest's unambiguous lead (0.250 accuracy) suggests that ensemble tree-based methods effectively capture the feature-label relationships in this classification task, even with randomly generated data showing slight deviation from pure chance. The three-way tie at 0.150 accuracy among Logistic Regression, Gradient Boosting, and Neural Network necessitates examination of secondary metrics for differentiation: Logistic Regression achieves the highest weighted F1-score (0.200) among the tied models, indicating better performance on the majority class; Gradient Boosting

Model	Accuracy	Precision	Recall	Macro -F1	Weighted -F1	ROC -AUC
Random Forest	0.2500	0.1583	0.3000	0.1587	0.2397	0.4386
Logistic Regression	0.1500	0.1500	0.0750	0.1000	0.2000	0.3938
Gradient Boosting	0.1500	0.2167	0.2500	0.1108	0.1677	0.5557
Neural Network (MLP)	0.1500	0.1200	0.0750	0.0923	0.1846	0.4567
SVM	0.0500	0.0111	0.2000	0.021	0.0053	0.6070

Table 4. Ranked Model Performance

demonstrates superior macro precision (0.217), reflecting more reliable positive predictions across classes; the Neural Network trails on both metrics, suggesting limited advantage from its non-linear architecture given the feature dimensionality. SVM's last-place ranking despite superior ROC-AUC represents a critical finding: the model's probability calibration or decision threshold is misaligned with the classification objective, warranting further investigation into threshold optimization or probability calibration techniques (e.g., Platt scaling or isotonic regression). The absence of XGBoost from the ranking due to library unavailability represents a limitation, as gradient boosted trees often outperform Random Forest on tabular data.

5.1 Reading the accuracy ranking

Random Forest leads at 0.25, the only classifier clearly above the others. Ranks 2–4 are a three-way tie at exactly 0.15. Logistic Regression, Gradient Boosting, and the Neural Network share an identical accuracy, so ranking them by accuracy alone is arbitrary; the tie-break comes from the secondary metrics (Logistic Regression has the highest weighted-F1 of the three at 0.2000, Gradient Boosting the highest precision at 0.2167, and the MLP trails on both). SVM is last at 0.05, the only model below its own precision floor; its confusion matrix shows predictions collapsing onto a single class despite the best ROC-AUC (0.6070), a sign the probability ranking carries signal, but the decision boundary is badly placed.

One caveat that governs how much weight to give this ordering: the script currently runs on its built-in dummy block ($X = \text{np.random.rand}(100, 9)$, $y = \text{np.random.randint}(0, 5, 100)$, $\text{seed}(42)$), i.e. random labels with no relationship to the features. With 5 classes, chance-level accuracy is 0.20 and no model reaches it except Random Forest, which barely grazes it. So, this ranking measures how these five model families split noise on a 20-row test set, not their comparative merit on real data. Swap in real features and labels and the ordering almost certainly reshuffles.

6. Discussion

6.1 Interpreting the Benchmark: Feature Sufficiency vs. Deep Learning

The primary goal of this benchmark is to determine if hand-crafted visual features are *sufficient* for style classification. If tree-based models (like Random Forest or Gradient Boosting) achieve high macro-F1 and ROC-AUC scores using only a 9-dimensional feature space, it mathematically proves that complex, computationally heavy deep learning architectures are largely unnecessary for this specific dimensional framing. This has massive implications for edge computing in smart home design applications, where lightweight,

interpretable models are preferred over resource heavy DNNs.

6.2 The SVM Paradox: Ranking vs. Hard Classification

A critical insight derived from the preliminary analytical framework is the divergent behavior of the SVM model. In scenarios where SVM exhibits poor hard-classification accuracy but superior ROC-AUC, it reveals a fundamental distinction in system design: probability ranking versus categorical tagging. For interior design applications, a user often benefits more from a ranked list of “style similarity scores” (e.g., “70% Scandinavian, 20% Modern, 10% Boho”) rather than a rigid, single label output. The SVM’s well calibrated probability estimates, despite a misaligned default decision threshold, make it highly valuable for recommendation engines, provided Platt scaling or isotonic regression is applied.

6.3 Addressing the “Black Box” through SHAP

By integrating SHAP values, this research moves beyond the “black box” of algorithmic design. If SHAP analysis confirms that features like *Mean Hue*, *Edge Density*, and *GLCM Contrast* account for the vast majority of predictive weight, it validates traditional interior design theories. For instance, it computationally proves that the “Industrial” style is defined not just by vague visual patterns, but by specific, quantifiable metrics of material roughness and linear structural rigidity. This bridges the gap between cognitive design perception and computational modeling.

6.4 Socio-Technical Implications and Dataset Bias

Interior Spatial Sensor Dataset inherently carries cultural assumptions about what constitutes “Boho,” “Modern,” or “Scandinavian” aesthetics. If the models heavily rely on specific color palettes (e.g., low-saturation whites for Scandinavian) to achieve high accuracy, it exposes a potential bias: the system may fail to recognize culturally diverse or evolving interpretations of these styles. Future iterations of this benchmark must incorporate fairness metrics to ensure that automated design tools do not enforce a narrow, Western-centric aesthetic standard.

7. Conclusion

Establishing a domain aware, comparative ML benchmark for interior design style classification is a vital step toward the standardization of automated design tools. By rigorously evaluating the sufficiency of engineered visual features and prioritizing model interpretability, this research provides a blueprint for developing intelligent systems that are not only accurate but also transparent, computationally efficient, and aligned with the functional and perceptual realities of human spatial experience. Transitioning from dummy data baselines to robust, cross validated real world testing will ultimately unlock reliable, context aware interior art placement and spatial optimization systems.

References

- [1] Grieze, E., Miélsone, E. (2021). Biomimicry element application in the interior design product development. *Economics. Ecology. Socium*, 5(2), 59–70.
- [2] Hasti, N., Kusnia, G. (2019, November). Role of technology for interior design sectors in creative economic development. In *IOP Conference Series: Materials Science and Engineering* (Vol. 662, No. 4). IOP Publishing.
- [3] Chen, Z., Wang, X. (2020). Application of AI technology in interior design. In *E3S Web of Conferences* (Vol. 179, p. 02105). E3S Web of Conferences.
- [4] Afifi, A. H. M. Y. (2020). Textile technology and its complementary design role in interior architecture. *International Journal of Architectural Engineering and Urban Research*, 3(1), 26–36.
- [5] Zhang, Z., Ban, J. (2022). Aesthetic evaluation of interior design based on visual features. *International Journal of Mobile Computing and Multimedia Communications (IJMCMC)*, 13(2), 1–12.

- [6] Cao, C. (2021, February). Research on innovative application of artistic characteristics of building decoration materials in interior design based on big data. *Journal of Physics: Conference Series*, 1744(2), 022020.
- [7] Zeng, J., Jiang, M. (2018, June). The role and application of decorative materials in interior design. *2nd International Conference on Social Sciences, Arts and Humanities (SSAH 2018)*, 9–10.
- [8] Viyanon, W., Songsuittipong, T., Piyapaisarn, P., Sudchid, S. (2017, July). AR furniture: Integrating augmented reality technology to enhance interior design using marker and markerless tracking. *Proceedings of the 2nd International Conference on Intelligent Information Processing*, 1–7.
- [9] Sarkar, A., Bardhan, R. (2019, July). A simulation based framework to optimize the interior design parameters for effective Indoor Environmental Quality (IEQ) experience in affordable residential units: Cases from Mumbai, India. In *IOP Conference Series: Earth and Environmental Science* (Vol. 294, No. 1, p. 012060). IOP Publishing.
- [10] Tong, H., Wan, Q., Kaszowska, A., Panetta, K., Taylor, H. A., Agaian, S. (2019). ARFurniture: Augmented reality interior decoration style colorization. *Electronic Imaging*, 2019(2), 175–1.
- [11] Jinkun, J., Yalan, L., Xizhu, Y. (2017). Application of Islamic decorative art in interior design. *Furniture & Interior Design*, 1(2), 8–19.
- [12] Wang, Z., Liu, D., Chang, S., Dolcos, F., Beck, D., Huang, T. (2017, May). Image aesthetics assessment using Deep Chatterjee's machine. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 941–948). IEEE.
- [13] Wang, W., Yang, S., Zhang, W., Zhang, J. (2019). Neural aesthetic image reviewer. *IET Computer Vision*, 13(8), 749–758.
- [14] Dou, Q., Zheng, X. S., Sun, T., Heng, P. A. (2019). Webthetics: Quantifying webpage aesthetics with deep learning. *International Journal of Human-Computer Studies*, 124, 56–66.
- [15] Jian, M., Zhou, Q., Cui, C., Nie, X., Luo, H., Zhao, J., Yin, Y. (2019). Assessment of feature fusion strategies in visual attention mechanism for saliency detection. *Pattern Recognition Letters*, 127, 37–47.
- [16] Zhang, S., Zhang, L. (2026). Automatic image composition optimization and visual perception simulation evaluation for architectural interior design. *Traitement du Signal*, 43(1), 505.
- [17] Fu, Y., Zhang, H. (2026, January). Color matching and style recommendation system for interior design based on computer vision and optical sensing. In *Second International Conference on Communication, Information, and Digital Technologies (CIDT 2025)* (Vol. 14064, p. 703–713). SPIE.
- [18] Hey, T., Butler, K., Jackson, S., Thiyaalingam, J. (2020). Machine learning and big scientific data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 378, 20190054.
- [19] Callaway, E. (2020). 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature*, 588, 203–204.
- [20] Takamoto, M., Praditia, T., Leiteritz, R., MacKinlay, D., Alesiani, F., Pflüger, D., Niepert, M. (2022). Pdebench: An extensive benchmark for scientific machine learning. *Advances in Neural Information Processing Systems*, 35, 1596–1611.
- [21] Raissi, M., Perdikaris, P., Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 686–707.

- [22] Greydanus, S., Dzamba, M., Yosinski, J. (2019). Hamiltonian neural networks. In H. Wallach et al. (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc.
- [23] Butler, K., Le, M., Thiyaalingam, J., Perring, T. (2021). Interpretable, calibrated neural networks for analysis and understanding of inelastic neutron scattering data. *Journal of Physics: Condensed Matter*, 33, 194006.
- [24] Thiyaalingam, J., Shankar, M., Fox, G., et al. (2022). Scientific machine learning benchmarks. *Nature Reviews Physics*, 4, 413–420. <https://doi.org/10.1038/s42254-022-00441-7>
- [25] Dongarra, J., Luszczek, P. (2011). Linpack. In D. Padua (Ed.), *Encyclopedia of Parallel Computing* (pp. 844–850). Springer.
- [26] Sakalis, C., Leonardsson, C., Kaxiras, S., Ros, A. (2016). Splash-3: A properly synchronized benchmark suite for contemporary research. In *2016 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)* (p. 101–111). IEEE.
- [27] Bailey, D. H. (2011). NAS parallel benchmarks. In D. Padua (Ed.), *Encyclopedia of Parallel Computing* (pp. 1254–1259). Springer.
- [28] Petitet, A., Whaley, R., Dongarra, J., Cleary, A. (2008). *HPL – A portable implementation of the high-performance Linpack benchmark for distributed-memory computers*. ICL-UTK Computer Science Department.
- [29] Dongarra, J., Luszczek, P. (2011). HPCG benchmark. In D. Padua (Ed.), *Encyclopedia of Parallel Computing* (pp. 2055–2057). Springer.
- [30] Denton, E. L., Hanna, A., Amironesei, R., et al. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, 8(2), 205395172110359.
- [31] Koch, B., Denton, E. L., Hanna, A., et al. (2021). Reduced, reused and recycled: The life of a dataset in machine learning research. *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, Sydney, Australia. <https://doi.org/10.48550/arXiv.2112.01716>.
- [32] Raji, I. D., Bender, E. M., Paullada, A., et al. (2022). AI and the everything in the whole wide world benchmark. *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA. arXiv:2111.15366: 1–20.
- [33] Mitchell, M., Luccioni, A. S., Lambert, N., et al. (2023). Measuring data. arXiv: 1–19. <https://doi.org/10.48550/arXiv.2212.05129>.
- [34] Engdahl, I. (2024). Agreements ‘in the wild’: Standards and alignment in machine learning benchmark dataset construction. *Big Data & Society*, 11(2). <https://doi.org/10.1177/20539517241242457>