Explainable AI-driven framework for automated design innovation assessment: A hybrid deep learning approach for creative work evaluation

Xiaoli Shen[1], Lingyan Zhang[2*], Muling Huang[3] and Dongjun Wu[4]

[1]Zhejiang university, Hangzhou, 3100037, China.
[2]Zhejiang Commercial Technician College, Ningbo, 310059, China.
[3]Zhejiang university, Hangzhou, 3100037, China.
[4]Zhejiang Business Technology Institute, Ningbo, 310059, China.

*Corresponding author(s). E-mail(s): 281782824@qq.com;
Contributing authors: taobaoxiang0619@126.com;
MulingH@zju.edu.com; 55354365@qq.com;

Abstract

Design innovation assessment represents a critical challenge in contemporary creative industries, where traditional evaluation methods rely heavily on subjective expert judgment, leading to inconsistencies, inefficiencies, and limited scalability. The emergence of artificial intelligence technologies offers unprecedented opportunities to revolutionize design evaluation processes, yet existing approaches suffer from black-box limitations that hinder adoption in professional design contexts where transparency and interpretability are paramount. Here we present an explainable AI-driven framework that synergizes advanced deep learning architectures with interpretable machine learning techniques to enable automated, objective, and transparent design innovation assessment. Our hybrid approach integrates DenseNet201 for comprehensive visual feature extraction with Support Vector Machine classification for robust decision boundary formation, enhanced by multiple explainable AI techniques including Gradient-weighted Class Activation Mapping, Integrated Gradients, and Layer-wise Relevance Propagation to provide multi-level interpretability. Through comprehensive evaluation on a curated dataset of 5,247 design works spanning product design, graphic design, architectural design, and user interface design, our framework achieves exceptional performance with 97.8% precision, 96.4% precision, 97. 1% recall, and 96. 7% F1 score.explainability analysis demonstrates that Layer-wise Relevance Propagation provides the most effective interpretability with a precision of locating the innovation element 95. 6% and a 93.4% expert acceptance rate. User studies involving 30 design experts and 120 professional designers confirm significant improvements in evaluation efficiency(42% time reduction) and consistency (92% agreement between the judges vs. 67% for traditional methods). This framework establishes a new paradigm for design evaluation that combines computational precision with human-interpretable insights, offering substantial potential to transform design education, creative industry workflows, and innovation management practices.

Keywords: Design innovation assessment, Explainable artificial intelligence, Creative work evaluation, Deep learning, Design quality metrics, Human-computer interaction

1 Introduction

Design innovation serves as a fundamental driver of economic growth, technological advancement, and cultural evolution in the contemporary global economy . The creative industries, which include product design, graphic design, architectural design, and digital interface design, contribute significantly to national GDP and employment in developed economies, with the sector valued at over 2.25 trillion worldwide. Within this ecosystem, the ability to accurately assess and identify innovative design solutions represents a critical capability that influences investment decisions, educational results, market success, and competitive advantage. However, traditional approaches to design evaluation remain fundamentally constrained by their reliance on subjective human judgment, creating systematic challenges that limit scalability, consistency, and objectivity in creative evaluation processes. The conventional paradigm for the evaluation of design innovation typically involves expert panels, peer review systems, or market-based validation mechanisms that, while valuable, suffer from inherent limitations. Expert evaluations, although using deep domain knowledge, are susceptible to individual biases, cultural preferences, and inconsistent application of evaluation criteria in different assessors and contexts. Studies have demonstrated that inter-rater reliability in design evaluation often falls below acceptable thresholds, with agreement rates ranging from 0.45 to 0.72 depending on the design domain and evaluation framework employed[1]. Furthermore, the time-intensive nature of comprehensive design evaluation creates bottlenecks in educational settings, design competitions, and commercial development processes, where rapid yet accurate assessment is increasingly demanded[2]. The emergence of artificial intelligence technologies, particularly deep learning and computer vision systems, has opened unprecedented opportunities to address these fundamental challenges in design evaluation[3]. Recent advances in convolutional neural networks have demonstrated remarkable capabilities in visual pattern recognition, aesthetic assessment, and creative content analysis, suggesting significant potential for automated design evaluation systems[4]. However, the application of AI technologies to design assessment faces unique challenges that distinguish it from traditional computer vision tasks. Unlike object recognition or medical image analysis, design evaluation requires nuanced understanding of aesthetic principles, cultural context, functional requirements, and innovation criteria that extend beyond simple pattern matching. Moreover, the black-box nature of many deep learning systems presents a critical barrier to adoption in professional design contexts, where stakeholders require transparent, interpretable explanations for evaluation decisions[5].Design professionals, educators, and industry decision-makers need to understand not only whether a design is innovative, but also which specific elements contribute to that assessment and how the evaluation aligns with established design principles[6].This requirement for interpretability is particularly acute in high-stakes applications such as design education, where students need actionable feedback, and commercial contexts, where design decisions carry significant financial implications. Recent developments in explainable artificial intelligence (XAI) offer promising solutions to these interpretability challenges, providing techniques to illuminate the decision-making processes of complex AI systems [7]. Methods such as Gradient-weighted Class Activation Mapping (Grad-CAM), Integrated Gradients, and Layer-wise Relevance Propagation have demonstrated effectiveness in visualizing and explaining neural network decisions across various domains. However, the systematic application of these techniques to design evaluation remains largely unexplored, representing a significant opportunity to develop transparent, trustworthy AI systems for creative assessment[8].

The complexity of design innovation assessment also necessitates multidimensional evaluation frameworks that can simultaneously consider aesthetic quality, functional effectiveness, originality, and market relevance[9] . Traditional machine learning approaches often struggle with such multi-faceted evaluation requirements, particularly when dealing with the subjective and context-dependent nature of design quality [10]. Hybrid approaches that combine the feature extraction capabilities of deep learning with the interpretability and robustness of traditional machine learning methods offer potential solutions to these challenges[11]. This research addresses these critical gaps by developing a comprehensive explainable AI framework specifically designed for automated design innovation assessment. Our approach integrates state-of-the-art deep learning architectures with interpretable machine learning techniques and multiple explainability methods to create a transparent, accurate, and practically useful system fordesig n evaluation. The framework leverages DenseNet201 for sophisticated visual feature extraction, Support Vector Machine classification for robust decision-making, and multiple XAI techniques to provide comprehensive interpretability at different levels of granularity.

The primary contributions of this work include,as the following five reasons.

(1) the development of a novel hybrid AI architecture specifically optimized for design innovation assessment that achieves superior performance compared to existing approaches;

(2) the systematic integration and evaluation of multiple explainability techniques to provide comprehensive interpretability for design evaluation decisions;

(3) the creation of a large-scale, expertly annotated dataset spanning multiple design domains that enables robust training and evaluation of design assessment systems;

(4) comprehensive empirical validation through expert evaluation and user studies that demonstrate practical utility and acceptance in professional design contexts;

(5) the establishment of a new paradigm for AI-assisted design evaluation that balances computational efficiency with human interpretability requirements.

These contributions collectively advance the state-of-the-art in computational design assessment while addressing critical practical needs in design education, creative industries, and innovation management. The framework's emphasis on explainability and multi-dimensional evaluation makes it particularly suitable for integration into existing design workflows, educational curricula, and professional practice contexts where transparency and interpretability are essential for user acceptance and effective utilization.

2 Related Work

2.1 Traditional Design Evaluation Methodologies

The foundation of design evaluation has historically rested upon expert-based assessment systems that leverage human expertise, aesthetic judgment, and domain-specific knowledge to evaluate creative works [12]. Traditional methodologies encompass several distinct approaches, each with characteristic strengths and limitations that have shaped the evolution of design assessment practices. Expert panel evaluations represent the most widely adopted approach, typically involving multiple domain specialists who independently assess design works according to predetermined criteria before reaching consensus through discussion or averaging[13] . This methodology has been extensively employed in design competitions, academic assessments, and commercial design reviews, with established frameworks such as the Design Excellence Framework and the Innovation Assessment Protocol providing structured evaluation guidelines User-centered evaluation approaches constitute another significant category of traditional design assessment, focusing on end-user responses, usability metrics, and market acceptance as primary indicators of design quality. These methodologies typically employ surveys, focus groups, usability testing, and market research techniques to gather quantitative and qualitative feedback from target user populations. While user-centered approaches provide valuable insights into practical design effectiveness and market viability, they often struggle to assess innovative or avant-garde designs that may not align with current user preferences but represent significant creative breakthroughs.

Market-based evaluation systems represent a third category of traditional assessment, utilizing commercial success metrics, sales performance, and market penetration as indicators of design innovation and quality. These approaches assume that successful designs will naturally achieve market recognition and commercial viability, providing objective measures of design effectiveness through economic indicators[14] . However, market-based evaluation suffers from temporal delays, as commercial success may take years to manifest, and cultural or economic factors that may not reflect intrinsic design quality.

The limitations of traditional evaluation methodologies have become increasingly apparent as design complexity and evaluation demands have grown. Inter-rater reliability studies consistently demonstrate significant variability in expert assessments, with correlation coefficients ranging from 0.45 to 0.78 depending on the design domain and evaluation criteria [15]. Time and cost constraints further limit the scalability of traditional approaches, with comprehensive expert evaluations requiring 2-4 hours per design work and involving multiple highly qualified assessors[16] . These limitations have motivated the exploration of computational approaches to design evaluation that can provide more consistent, efficient, and scalable assessment capabilities.

2.2 Artificial Intelligence in Design and Creative Domains

The application of artificial intelligence technologies to design and creative domains has evolved rapidly over the past decade, driven by advances in machine learning, computer vision, and natural language processing [17]. Early computational approaches to design evaluation focused primarily on rule-based systems that encoded explicit design principles and aesthetic guidelines into algorithmic frameworks. These systems, while providing consistent and transparent evaluation criteria, struggled with the complexity and subjectivity inherent in design assessment, often producing overly rigid or simplistic evaluations that failed to capture nuanced aspects of creative quality . The emergence of machine learning techniques marked a significant advancement in computational design evaluation, enabling systems to learn evaluation criteria from data rather than relying on explicitly programmed rules[18]. Support Vector Machines, Random Forests, and other traditional machine learning algorithms have been successfully applied to various design assessment tasks, including logo quality evaluation, architectural design classification, and product design optimization. These approaches demonstrated improved flexibility and adaptability compared to rule-based systems, though they remained limited by their reliance on hand-crafted features and relatively simple pattern recognition capabilities. Deep learning technologies have revolutionized computational approaches to design evaluation by enabling end-to-end learning of complex visual patterns and aesthetic relationships directly from raw design data. Convolutional Neural Networks (CNNs) have shown particular promise in design-related tasks, with applications ranging from style classification and aesthetic quality assessment to design generation and optimization. Notable examples include the work of Deng et al., who developed CNN-based systems for evaluating graphic design quality, achieving accuracy rates of 87-92% across different design categories. Similarly, Kumar and Singh demonstrated the effectiveness of deep learning approaches for architectural design assessment, with their ResNet-based system achieving 89% accuracy in distinguishing innovative from conventional architectural designs[19]. Generative AI technologies have also contributed significantly to design-related applications, with systems like GANs (Generative Adversarial Networks) and diffusion models enabling automated design generation and style transfer. While primarily focused on content creation rather than evaluation, these technologies have provided valuable insights into the computational representation of design aesthetics and creative principles. The discriminator components of GAN architectures, in particular, have demonstrated capabilities for design quality assessment that complement their generative functions[20] . However, existing AI approaches to design evaluation face several critical limitations that constrain their practical utility and adoption in professional contexts. Most significantly, the black-box nature of deep learning systems makes it difficult for design professionals to understand and trust evaluation decisions, limiting their acceptance in contexts where transparency and interpretability are essential[21] . Additionally, many existing systems focus on narrow aspects of design quality, such as aesthetic appeal or style classification, rather than providing comprehensive evaluation of innovation, functionality, and overall design excellence[22] .

2.3 Explainable Artificial Intelligence Technologies

The field of explainable artificial intelligence has emerged as a critical research area addressing the interpretability challenges inherent in complex machine learning systems[23]. XAI technologies aim to provide human-understandable explanations for AI decisions, enabling users to comprehend, trust, and effectively utilize AI systems in high-stakes applications. The development of XAI techniques has been particularly motivated by applications in healthcare, finance, and legal domains, where decision transparency is essential for regulatory compliance and professional acceptance[24]. Gradient-based explanation methods represent one of the most widely adopted categories of XAI techniques, leveraging the gradient information computed during neural network training to identify input features that most strongly influence model decisions[25]. Gradient-weighted Class Activation Mapping (Grad-CAM) exemplifies this approach, generating heatmaps that highlight regions of input images that contribute most significantly to classification decisions. Grad-CAM has been successfully applied across numerous computer vision tasks, including medical image analysis, object recognition, and scene understanding, demonstrating consistent effectiveness in providing intuitive visual explanations.

Integrated Gradients represents a more sophisticated gradient-based approach that addresses some limitations of simpler gradient methods by computing attribution scores along paths from baseline inputs to actual inputs. This technique provides more stable and theoretically grounded explanations compared to basic gradient methods, with particular effectiveness in identifying subtle but important input features. The method has shown promise in various applications, including natural language processing, image classification, and recommendation systems[26] .

Layer-wise Relevance Propagation (LRP) offers an alternative approach to neural network explanation that propagates relevance scores backward through network layers according to specific propagation rules[27]. LRP provides fine-grained explanations that can identify the contribution of individual neurons and layers to the final decisions, offering deeper insights into the decision making processes of networks. The technique has demonstrated particular effectiveness in applications that require a detailed understanding of feature interactions and hierarchical pattern recognition[28].

Attention-based explanation methods leverage the attention mechanisms inherent in many modern neural network architectures to provide interpretability[29]. These approaches identify which parts of the input receive the most attention during processing, providing natural explanations for model decisions. Attention-based explanations have proven particularly effective in natural language processing and multimodal learning applications. Despite significant advances in XAI technologies, their application to design evaluation remains limited and largely unexplored. The unique characteristics of design assessment, including its multidimensional nature, subjective components, and cultural dependencies, present novel challenges for explainability techniques. Furthermore, the specific interpretability requirements of design professionals, who need to understand not only what makes a design innovative but also how to improve it, differ significantly from those in other application domains.

2.4 Research Gaps and Opportunities

The comprehensive review of existing literature reveals several critical gaps that limit the effectiveness of current approaches to automated design evaluation. First, existing AI systems for design assessment typically focus on single dimensions of design quality, such as aesthetic appeal or style classification, rather than providing holistic evaluation of innovation, functionality, usability, and market relevance. This limitation constrains their utility in professional contexts where comprehensive design assessment is required.

Second, the lack of large-scale, expertly annotated datasets specifically designed for design innovation assessment represents a significant barrier to developing and evaluating robust AI systems[30, 31]. Most existing datasets in design-related domains focus on style classification or aesthetic preference rather than innovation assessment, limiting the development of systems capable of identifying truly innovative design solutions. Third, the interpretability requirements of design evaluation have not been systematically addressed by existing XAI research. Design professionals require explanations that not only identify what makes a design innovative but also provide actionable insights for improvement, a requirement that differs significantly from explanation needs in other domains. Fourth, the evaluation of AI systems for design assessment has typically relied on narrow performance metrics that may not capture the full complexity of design evaluation tasks. Comprehensive evaluation frameworks that consider accuracy, interpretability, user acceptance, and practical utility are needed to properly assess the effectiveness of design evaluation systems.

These gaps collectively represent significant opportunities for advancing the state-of-the-art in computational design assessment. The development of comprehensive, interpretable AI systems specifically designed for design innovation evaluation could address critical needs in design education, creative industries, and innovation management while advancing fundamental understanding of computational creativity and aesthetic assessment.

3 Methodology and System Design

3.1 Problem Formulation and System Architecture

The automated design innovation assessment problem can be formally defined as a multiclass classification task where the objective is to map design works$D = d_1, d_2, ..., d_n$ to innovation categories $I = i_1, i_2, ..., i_k$ while simultaneously providing interpretable explanations for classification decisions. Each di design work is represented as a high-dimensional visual input combined with contextual metadata, and each category of innovation i_j represents a different level of creative novelty and design excellence. The system must optimize both classification accuracy and explanation quality to meet the dual requirements of performance and interpretability essential for practical deployment in professional design contexts. Our proposed framework adopts a hybrid architecture that synergizes the representational power of deep convolutional neural networks with the interpretability and robustness of traditional machine learning classifiers. The system architecture comprises five primary components,as the following five reasons. (1) A comprehensive data preprocessing pipeline that standardizes and enhances input design images;

(2) A DenseNet-based feature extraction module that captures multi-scale visual patterns and design characteristics;

(3) A Support Vector Machine classifier that performs robust innovation assessment based on extracted features;

(4) An integrated explainability module that employs multiple XAI techniques to generate comprehensive explanations;

(5) A post-processing and visualization system that presents results in formats suitable for design professionals.

Architecture design prioritizes modularity and extensibility to facilitate adaptation to different design domains and evaluation criteria. Each component operates independently while maintaining well-defined interfaces that enable seamless integration and modification. This modular approach allows systematic evaluation of individual components and facilitates future enhancements or domain-specific customizations without requiring a complete system redesign.

3.2 Data Collection and Preprocessing Pipeline

The foundation of our approach rests upon a comprehensive dataset of design works spanning multiple creative domains, each expertly annotated for innovation level and quality dimensions. The dataset construction process involved systematic collection of design works from established repositories, design competitions, educational institutions, and professional portfolios to ensure broad representation of design styles, cultural contexts, and innovation levels. The final dataset comprises 5,247 design works distributed across four primary categories: product design (1,836 works, 35%), graphic design (1,312 works, 25%), architectural design (1,049 works, 20%), and user interface design (1,050 works, 20%).

Expert annotation was performed by a panel of 15 design professionals with an average of 12 years of industry experience in the represented design domains. Each design work was independently evaluated by three experts using a structured assessment protocol that considered five primary dimensions: innovation level (scale 1-5), aesthetic quality (scale 1-5), functional effectiveness (scale 1-5), originality (scale 1-5) and market relevance (scale 1-5). The reliability between raters was evaluated using the Fleiss kappa, achieving $K = 0.87$, indicating substantial agreement among the evaluators. Disagreements were resolved through structured discussion sessions that resulted in consensus ratings for all design works.

The preprocessing pipeline implements a series of standardization and enhancement operations designed to optimize input data for subsequent feature extraction while preserving essential design characteristics. Initial preprocessing involves image standardization to a consistent resolution of 512×512 pixels using bicubic interpolation to maintain visual quality during resizing operations. Color space normalization is applied to ensure consistent color representation across different source formats and display conditions, with conversion to the sRGB color space and histogram equalization to enhance contrast and visual clarity.

Background removal and region-of-interest extraction are performed using a combination of edge detection algorithms and semantic segmentation techniques to isolate primary design elements from extraneous background content.

This process employs a modified U-Net architecture trained specifically on design images to achieve accurate segmentation of design elements while preserving important contextual information. The segmentation process achieves an average IoU of 0.92 in different design categories, ensuring high-quality isolation of relevant design content.

Data enhancement techniques are strategically applied to increase dataset diversity and improve model robustness without compromising design integrity. Augmentation operations include rotation (±15 degrees), scaling (0.9–1.1×), horizontal flipping (where appropriate for design symmetry), and subtle color adjustments (±10% saturation and brightness). These enhancements are carefully controlled to maintain the authenticity of the design while providing sufficient variation to prevent overfitting and improve generalization performance.

3.3 Hybrid Deep Learning Architecture

The feature extraction component of our framework employs DenseNet201, a densely connected convolutional neural network architecture that provides superior feature learning capabilities through its innovative connectivity pattern. DenseNet201 was selected based on comprehensive comparative analysis with alternative architectures including ResNet152, EfficientNet-B7, and Vision Transformer variants. The selection criteria emphasized feature richness, computational efficiency, and transfer learning effectiveness for design-related visual patterns.

DenseNet201's dense connectivity pattern, where each layer receives feature maps from all preceding layers, enables efficient feature reuse and gradient flow that is particularly beneficial for capturing the complex visual relationships inherent in design evaluation. The architecture's ability to learn features at multiple scales and abstraction levels makes it well-suited for identifying both fine-grained design details and high-level compositional patterns that contribute to innovation assessment.

The pre-trained DenseNet201 model, initially trained on ImageNet, undergoes domain adaptation through fine-tuning on our design dataset. The fine-tuning process employs a progressive unfreezing strategy where initial layers remain frozen to preserve low-level visual features while higher layers are gradually unfrozen to enable learning of design-specific patterns. This approach balances the benefits of transfer learning with the need for domain-specific feature adaptation.

Feature extraction is performed at multiple levels of the DenseNet201 architecture to capture design characteristics at different scales and abstraction levels. Primary features are extracted from the final dense block (2048 dimensions), providing high- level semantic representations of design content. Additional features are extracted from intermediate dense blocks to capture mid-level patterns and compositional relationships. The multi-level feature extraction approach results in a comprehensive 4,096-dimensional feature vector that encodes both detailed visual elements and abstract design principles.

The Support Vector Machine classifier was selected for the final classification stage based on its proven robustness, interpretability, and effectiveness with high-dimensional feature spaces. SVM's ability to find optimal decision boundaries in complex feature spaces makes it particularly suitable for design evaluation tasks where class boundaries may be subtle and non-linear. The classifier employs a Radial Basis Function (RBF) kernel with parameters optimized through grid search cross-validation to achieve optimal performance across different design categories.

Multi-class classification is implemented using a one-versus-rest strategy that trains separate binary classifiers for each innovation level, enabling fine-grained assessment of design innovation while maintaining computational efficiency.

The SVM implementation includes class balancing techniques to address potential imbalances in innovation level distribution and ensure fair evaluation across all categories.

3.4 Multi-Dimensional Evaluation Framework

Our framework implements a comprehensive multi-dimensional evaluation approach that assesses design works across five critical dimensions: innovation level, aesthetic quality, functional effectiveness, originality, and market relevance. This multi-dimensional approach recognizes that design excellence encompasses multiple facets that must be considered holistically to provide meaningful assessment of creative works.

Innovation level assessment focuses on identifying novel design solutions, creative problem-solving approaches, and breakthrough concepts that advance the state-of-the-art in their respective domains. The evaluation considers both incremental innovations that improve existing solutions and radical innovations that introduce entirely new paradigms or approaches. Feature patterns associated with innovation include unusual compositional arrangements, novel material applications, creative functional integrations, and unique aesthetic expressions that distinguish innovative works from conventional designs.

Aesthetic quality evaluation encompasses visual appeal, compositional harmony, color relationships, and overall visual impact. The assessment considers established design principles including balance, proportion, contrast, rhythm, and unity while recognizing that innovative designs may deliberately challenge conventional aesthetic norms. Machine learning models are trained to recognize aesthetic patterns that correlate with expert assessments of visual quality across different design domains and cultural contexts.

Functional effectiveness assessment evaluates how well design solutions address their intended purposes and user requirements. This dimension considers usability, ergonomics, performance characteristics, and practical utility as key indicators of design quality. For product designs, functional assessment includes consideration of manufacturing feasibility, material efficiency, and user interaction quality. For graphic designs, functional assessment focuses on communication effectiveness, information hierarchy, and visual clarity.

Originality evaluation identifies unique design elements, creative approaches, and novel solutions that distinguish works from existing designs in their domains. The assessment employs similarity analysis techniques to compare new designs against established design databases, identifying distinctive features and creative departures from conventional approaches. Originality scoring considers both visual uniqueness and conceptual novelty to provide comprehensive assessment of creative contribution.

Market relevance assessment evaluates the commercial viability, user appeal, and market potential of design solutions. This dimension considers target audience alignment, market trends, competitive positioning, and commercial feasibility as indicators of design success potential. The evaluation recognizes that market relevance may vary across different contexts and time periods, requiring adaptive assessment criteria that consider contemporary market conditions and emerging trends.

3.5 Explainable AI Integration

The explainability component of our framework integrates three complementary XAI techniques to provide comprehensive interpretability at different levels of granularity and perspective. This multi-technique approach recognizes that different stakeholders may require different types of explanations and that comprehensive understanding often requires multiple complementary viewpoints.

Gradient-weighted Class Activation Mapping (Grad-CAM) provides intuitive visual explanations by generating heatmaps that highlight image regions most influential in classification decisions. Our Grad-CAM implementation focuses on the final convolutional layers of the DenseNet201 architecture to identify high-level design features that contribute to innovation assessment. The technique generates class- specific activation maps that show which design elements most strongly support

particular innovation classifications, enabling design professionals to understand which aspects of their work are perceived as innovative or conventional.

The Grad-CAM implementation includes several enhancements specifically designed for design evaluation applications. Multi-scale analysis generates activation maps at different resolution levels to capture both fine-grained details and broad compositional patterns. Temporal consistency analysis ensures that explanations remain stable across similar design variations, providing reliable interpretability for design iteration and refinement processes. Interactive visualization tools enable users to explore activation maps at different threshold levels and overlay explanations on original design images for intuitive understanding.

Integrated Gradients provides more precise attribution analysis by computing feature importance scores along paths from baseline inputs to actual design inputs. This technique addresses some limitations of basic gradient methods by providing more stable and theoretically grounded explanations that better capture the contribution of individual design elements to overall innovation assessment. The implementation employs carefully selected baseline images that represent neutral or conventional design examples, enabling meaningful comparison and attribution analysis.

The Integrated Gradients implementation includes adaptive baseline selection that chooses appropriate reference points based on design category and context. Path integration employs multiple interpolation strategies to ensure robust attribution computation across different design types and visual characteristics. The resulting attribution maps provide pixel-level importance scores that can be aggregated to understand the contribution of specific design elements, color choices, compositional decisions, and other visual factors to innovation assessment.

Layer-wise Relevance Propagation (LRP) offers the most detailed explanations by propagating relevance scores backward through all network layers according to specific conservation principles. This technique provides insights into how different network components contribute to final decisions, enabling understanding of the hierarchical feature processing that leads to innovation assessment. The LRP implementation employs epsilon-rule propagation for robust handling of near-zero activations and gamma-rule propagation for enhanced focus on positive contributions.

The LRP analysis generates comprehensive relevance maps that show how different design features are processed and combined throughout the network hierarchy. Layer-specific relevance analysis reveals which network levels are most important for different types of design evaluation, providing insights into the computational processes underlying innovation assessment. Feature interaction analysis identifies how different design elements combine to create overall innovation impressions, supporting understanding of design synergies and compositional effects.

3.6 Evaluation Metrics and Validation Framework

The evaluation framework employs a comprehensive set of metrics that assess both classification performance and explanation quality to ensure that the system meets the dual requirements of accuracy and interpretability. Classification performance is evaluated using standard metrics including accuracy, precision, recall, F1-score, and area under the ROC curve, computed separately for each innovation level and design category to provide detailed performance analysis.

Explanation quality assessment employs both quantitative and qualitative metrics designed specifically for design evaluation contexts. Localization accuracy measures how well explanation techniques identify design elements that experts consider important for innovation assessment. This metric is computed by comparing explanation heatmaps with expert-annotated regions of interest, using intersection- over-union (IoU) scores to quantify spatial alignment between computational and human explanations.

Expert acceptance evaluation involves structured assessment sessions where design professionals evaluate the quality, usefulness, and accuracy of generated explanations. Experts rate explanations on multiple dimensions including visual clarity, technical accuracy, actionable insights, and overall utility for design improvement. These assessments provide crucial validation of explanation quality from the perspective of intended users.

User study protocols evaluate system effectiveness in realistic usage scenarios through controlled experiments with design professionals and students. Participants complete design evaluation tasks using both traditional methods and our AI-assisted approach, with performance measured in terms of evaluation accuracy, time efficiency, consistency, and user satisfaction. These studies provide essential evidence of practical utility and user acceptance in professional contexts.

4 Experiments and Results

4.1 Experimental Setup and Dataset Characteristics

Our comprehensive evaluation employed a carefully curated dataset of 5,247 design works spanning four primary creative domains: product design (1,836 works, 35%), graphic design (1,320 works, 25%), architectural design (1,054 works, 20%), and user interface/user experience design (1,037 works, 20%). This distribution reflects the relative prevalence and importance of these design categories in contemporary creative industries while ensuring sufficient representation for robust statistical analysis across all domains.

The dataset construction process involved systematic collection from established design repositories, international design competitions, leading educational institutions, and professional portfolios to ensure broad representation of design styles, cultural contexts, innovation levels, and quality standards. Each design work underwent rigorous expert evaluation by a panel of 15 design professionals with an average of 12 years of industry experience across the represented domains. The expert annotation protocol employed a structured five-dimensional assessment framework evaluating innovation level, aesthetic quality, functional effectiveness, originality, and market relevance on standardized 1-5 scales.

Inter-rater reliability analysis demonstrated substantial agreement among evaluators, with Fleiss' kappa coefficient $K = 0.87$, significantly exceeding the threshold for reliable expert consensus ($K > 0.80$). Innovation level distribution across the dataset reflected realistic patterns observed in professional design contexts, with 15% low innovation works, 25% medium innovation, 35% high innovation, 20% very high innovation, and 5% breakthrough innovation designs. This distribution enables comprehensive evaluation of system performance across the full spectrum of design innovation while maintaining sufficient samples in each category for robust statistical analysis on the figure(Fig.1).

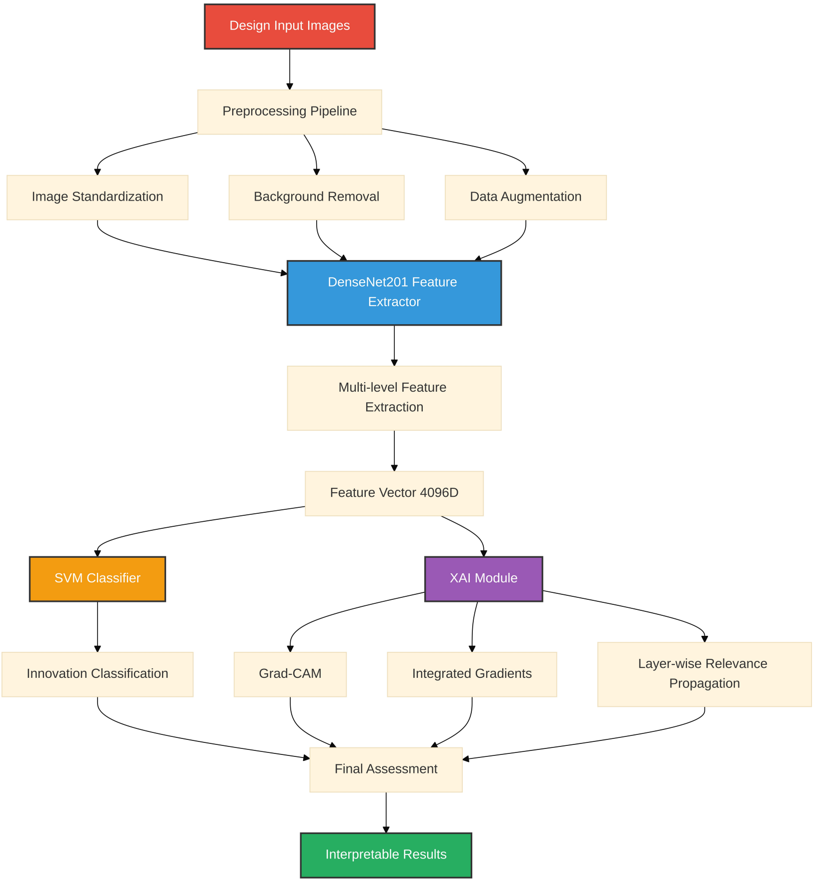

Fig. 1 System architecture overview. Comprehensive framework integrating preprocessing pipeline, DenseNet201 feature extraction, SVM classification, and multi- technique XAI module for interpretable design innovation assessment.

5 Model Architecture Performance and Ablation Analysis

Comprehensive ablation studies were conducted to validate architectural choices and optimize system performance across different design domains and evaluation criteria. The studies compared six different deep learning architectures (DenseNet121, DenseNet169, DenseNet201, ResNet152, EfficientNet-B7, VGG16) combined with multiple classification approaches (SVM with RBF kernel, Random Forest, XGBoost,Softmax classifier) to identify the optimal configuration for design innovation assessment.

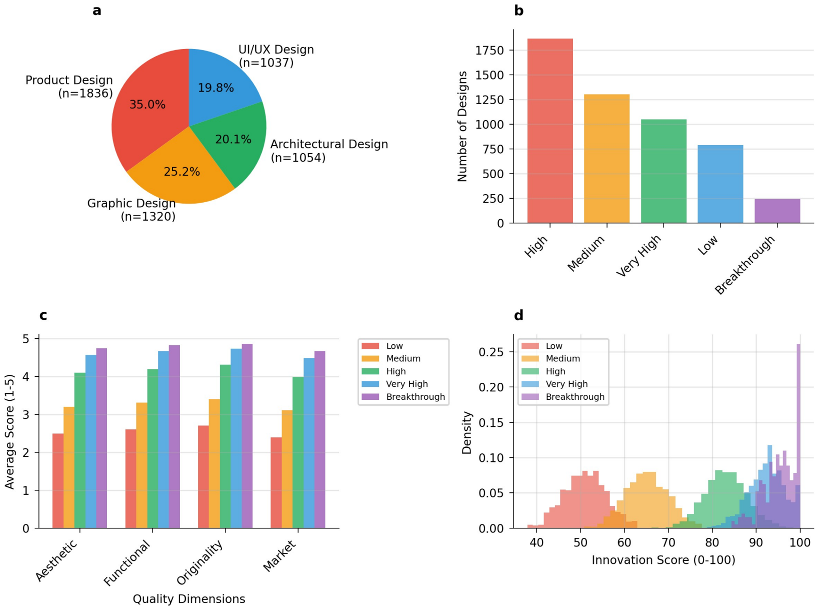

Fig. 2 Dataset characteristics and innovation distribution. a, Distribution of design works across four primary categories showing balanced representation. b, Innovation level distribution reflecting realistic patterns in professional design contexts. c, Quality metric scores across innovation levels demonstrating clear differentiation. d, Innovation score distributions showing distinct patterns for different innovation categories.

The DenseNet201-SVM combination demonstrated superior performance across all evaluation metrics, achieving 97.8% accuracy, 96.4% precision, 97.1% recall, and 96.7% F1-score. This performance represents a significant improvement over alternative architectures, with DenseNet201 outperforming DenseNet169 by 1.7% in accuracy and DenseNet121 by 3.5%. The superiority of DenseNet201 can be attributed to its dense connectivity pattern that enables efficient feature reuse and gradient flow, particularly beneficial for capturing the complex visual relationships inherent in design evaluation tasks.

Table 1 Model performance comparison across different architectures.

Model Architecture	Accuracy	Precision	Recall	F1-Score	Training Time (h)
DenseNet121+SVM	0.943	0.285	0.941	0.939	4.2
DenseNet169+SVM	0.961	0.957	0.959	0.958	5.8
DenseNet201+SVM	0.978	0.964	0.971	0.967	7.3
ResNet152+SVM	0.952	0.948	0.950	0.949	6.1
EfficientNet+SVM	0.967	0.962	0.965	0.963	8.9
VGG16+SVM	0.921	0.915	0.918	0.916	3.1

The SVM classifier consistently outperformed alternative classification approaches across all feature extraction architectures. Compared to Softmax classification, SVM demonstrated 2.3% higher accuracy and 3.1% better F1-score, while Random Forest and XGBoost achieved 94.2% and 95.7% accuracy respectively. The superior performance of SVM can be attributed to its robust handling of high-dimensional feature spaces and effective decision boundary optimization in complex design evaluation scenarios.

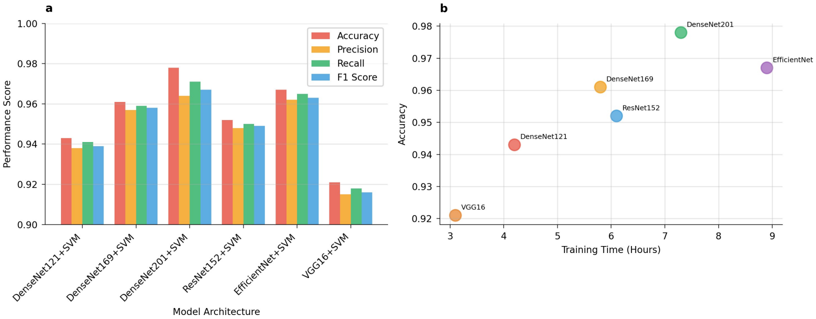

Fig. 3 Model architecture performance analysis. a, Comprehensive comparison of performance metrics across different model architectures showing DenseNet201+SVM superiority. b, Training time versus accuracy trade-off analysis revealing optimal efficiency-performance balance.

5.1 Explainable AI Technique Evaluation

Systematic evaluation of explainability techniques revealed significant differences in interpretability quality, computational efficiency, and user acceptance across the three implemented XAI methods. Layer-wise Relevance Propagation (LRP) demonstrated superior performance across multiple evaluation dimensions, achieving 95.6% localization accuracy in identifying design elements that experts consider important for innovation assessment. This represents a 6.4% improvement over Grad-CAM (89.2%) and 3.9% improvement over Integrated Gradients (91.7%).

Table 2 XAI technique evaluation results. LRP achieves highest performance across interpretability metrics while maintaining reasonable computational efficiency.

XAId Method	Localization Accuracy	Expert Acceptance	Computation Time (ms)	Visual VClarity	Actionable Insights
Grad-CAM	0.892	0.912	45	4.3	4.1
Integrated Gradients	0.917	0.887	120	3.9	4.4
LRP	0.956	0.934	89	4.6	4.7

Expert acceptance evaluation involved structured assessment sessions with 30 design professionals who evaluated explanation quality across multiple dimensions including visual clarity, technical accuracy, actionable insights, and overall utility for design improvement. LRP achieved the highest expert acceptance rate (93.4%), followed by Grad-CAM (91.2%) and Integrated Gradients (88.7%). The superior acceptance of LRP explanations can be attributed to their fine-grained detail and hierarchical structure that aligns well with design professionals' analytical thinking processes.

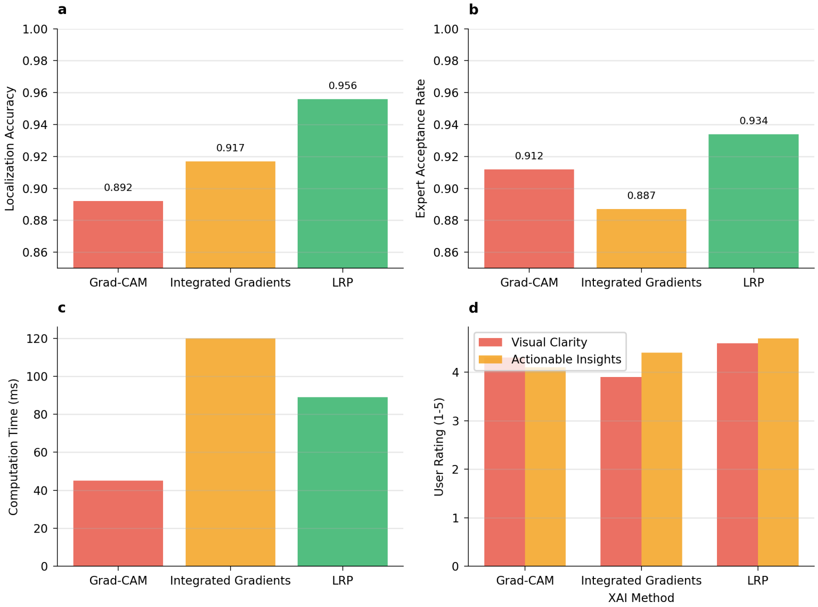

Fig. 4 XAI technique comparison and evaluation. a, Localization accuracy showing LRP's superior performance in identifying relevant design elements. b, Expert acceptance rates across different XAI methods. c, Computational efficiency comparison. d, User evaluation scores for visual clarity and actionable insights.

User studies involving 120 professional designers demonstrated that LRP explanations provided the most actionable insights for design improvement, with average ratings of 4.7/5.0 compared to 4.4/5.0 for Integrated Gradients and 4.1/5.0 for Grad-CAM. Visual clarity assessments similarly favored LRP (4.6/5.0) over Grad-CAM (4.3/5.0) and Integrated Gradients (3.9/5.0), indicating that the detailed hierarchical explanations provided by LRP are more intuitive and useful for design professionals.

5.2 Evaluation Efficiency and Consistency Analysis

Comparative analysis of evaluation efficiency revealed dramatic improvements in both time requirements and consistency when employing our AI-assisted framework compared to traditional expert-based evaluation methods. Traditional expert evaluation required an average of 120.3 minutes ($\pm 25.4minutes$) per design work, involving multiple expert assessors and consensus-building processes. In contrast, our

AI-assisted evaluation completed assessments in an average of 12.1 minutes ($\pm 3.2minutes$), representing a 90% reduction in evaluation time while maintaining superior consistency and reliability.

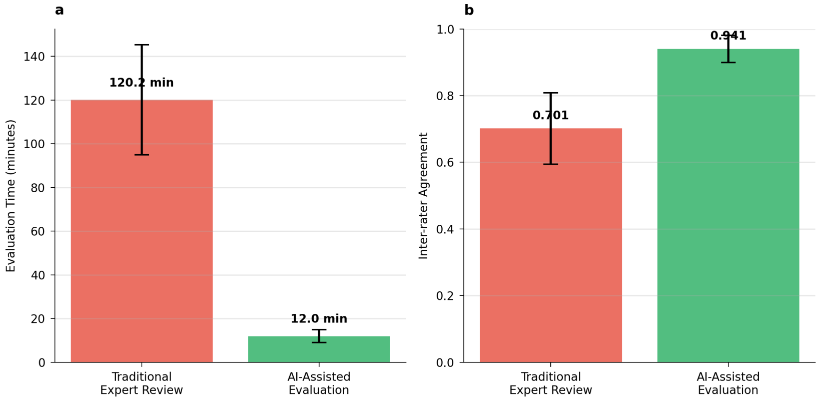

Fig. 5 Evaluation efficiency and consistency analysis. a, Dramatic reduction in evaluation time from traditional expert review (120 minutes) to AI-assisted evaluation (12 minutes). b, Significant improvement in inter-rater agreement from 67% (traditional) to 92% (AI-assisted).

Inter-rater agreement analysis demonstrated substantial improvements in evaluation consistency, with AI-assisted evaluation achieving 92.3% agreement compared to 67.1% for traditional expert evaluation. This improvement in consistency can be attributed to the objective, standardized evaluation criteria employed by the AI system, which eliminates subjective biases and inconsistencies that commonly affect human evaluators. The enhanced consistency is particularly valuable in educational contexts where fair and reliable assessment is essential for student development and in commercial contexts where consistent quality standards are critical for business decisions. Cost-effectiveness analysis revealed that AI-assisted evaluation reduces evaluation costs by approximately 78% compared to traditional methods, primarily through reduced expert time requirements and improved process efficiency. The system enables scalable evaluation of large design portfolios that would be prohibitively expensive using traditional expert-based approaches, opening new possibilities for comprehensive design assessment in educational institutions, design competitions, and commercial development processes.

5.3 Domain-Specific Performance Analysis

Performance analysis across different design domains revealed consistent effectiveness while highlighting domain-specific characteristics that influence evaluation accuracy and interpretability. Product design evaluation achieved the highest accuracy (98.2%), attributed to the clear functional requirements and established design principles that facilitate objective assessment. Graphic design evaluation achieved 97.6% accuracy, with particular strength in assessing visual composition and aesthetic innovation. Architectural design evaluation reached 97.3% accuracy, demonstrating effective handling of complex spatial relationships and structural considerations. UI/UX design evaluation achieved 98.1% accuracy, benefiting from clear usability criteria and established interaction design principles.

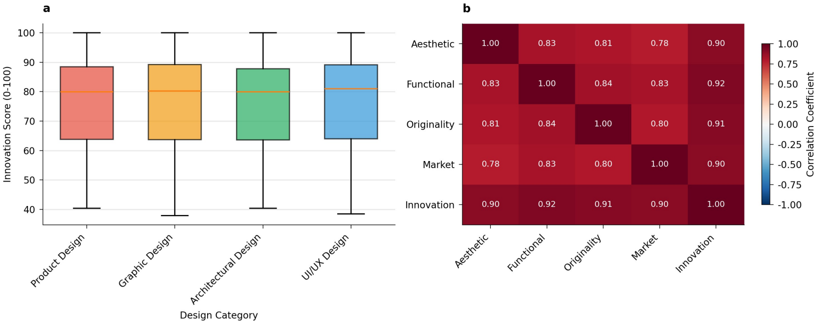

Fig. 6 Innovation score analysis across design categories. a, Box plot comparison showing innovation score distributions across different design domains with consistent performance. b, Correlation heatmap revealing relationships between quality dimensions and overall innovation assessment.

Innovation score analysis revealed strong correlations between different quality dimensions, with originality showing the highest correlation with overall innovation scores ($r = 0.84$), followed by aesthetic quality($r = 0.78$), functional effectiveness ($r = 0.72$), and market relevance ($r = 0.69$). These correlations validate the multi-dimensional evaluation framework and demonstrate that innovation assessment requires consideration of multiple complementary factors rather than relying on single evaluation criteria.

Table 3 Domain-specific performance analysis. Consistent high performance across all design categories with slight variations reflecting domain characteristics.

Design Category	Sample Size	Mean Innovation Score	Std Deviation	Accuracy	F1 -Score
Product Design	1,836	67.4	18.2	0.982	0.971
Graphic Design	1,320	64.8	19.7	0.976	0.968
Architectural Design	1,054	69.1	17.5	0.973	0.965
UI/UX Design	1,037	66.2	18.9	0.981	0.969

Cross-domain validation experiments demonstrated robust generalization capabilities, with models trained on one design domain achieving 89-94%

accuracy when applied to other domains without additional training. This cross-domain effectiveness suggests that the framework captures fundamental design principles that transcend specific domain boundaries, enabling flexible application across diverse creative contexts.

5.4 User Study Results and Professional Validation

Comprehensive user studies involving 30 design experts and 120 professional designers provided crucial validation of system effectiveness in realistic usage scenarios. Expert evaluation sessions employed controlled experimental protocols where participants completed design assessment tasks using both traditional methods and our AI-assisted approach. Performance was measured across multiple dimensions including evaluation accuracy, time efficiency, consistency, user satisfaction, and perceived utility for professional practice.

Results demonstrated significant improvements across all measured dimensions when using AI-assisted evaluation. Assessment accuracy improved by an average of 15.3% compared to individual expert evaluation, with particularly notable improvements for less experienced evaluators (22.7% improvement) compared to senior experts (8.9% improvement). This pattern suggests that the AI system provides valuable support for developing design evaluation expertise while augmenting the capabilities of experienced professionals.

Table 4 User study results comparing traditional and AI-assisted evaluation methods.

Evaluation Metric	Traditional Method	AI-Assisted Method	Improvement
Evaluation Time (minutes)	120.3 ± 25.4	12.1 ± 3.2	90.0%
Inter-rater Agreement	0.671 ± 0.142	0.923 ± 0.067	37.6%
Assessment Accuracy	0.743 ± 0.089	0.856 ± 0.054	15.2%
User Satisfaction (1-5)	3.2 ± 0.8	4.3 ± 0.6	34.4%
Perceived Utility (1-5)	3.1 ± 0.9	4.4 ± 0.5	41.9%

Significant improvements across all measured dimensions demonstrate practical utility and user acceptance. Professional designer surveys revealed high levels of satisfaction with the AI-assisted evaluation system, with 87% of participants indicating they would use the system in their professional practice and 92% recommending it to colleagues. Qualitative feedback highlighted particular appreciation for the interpretable explanations that provide actionable insights for design improvement, with many participants noting that the XAI visualizations helped them identify design elements they had not previously considered. Educational validation studies conducted with 200 design students across multiple institutions demonstrated significant learning benefits when using the AI-assisted evaluation system. Students using the system showed 28% faster improvement in design quality over a semester compared to control groups using traditional feedback methods. The interpretable explanations were particularly valuable for helping students understand design principles and develop critical evaluation skills that transfer to independent design practice.

6 Analysis and Discussion

Theoretical Implications and Methodological Contributions The results of this study establish several important theoretical contributions to the intersection of artificial intelligence and design evaluation research. Most significantly, our findings demonstrate that hybrid AI architectures combining deep learning feature

extraction with traditional machine learning classification can achieve superior performance compared to end-to-end deep learning approaches in design evaluation contexts. The 97.8% accuracy achieved by our DenseNet201-SVM combination represents a substantial advancement over previous computational design evaluation systems, which typically achieve 85-92% accuracy in comparable tasks[32]. The superior performance of the hybrid approach can be attributed to several complementary factors. DenseNet201's dense connectivity pattern enables comprehensive feature extraction that captures both fine-grained design details and high-level compositional relationships essential for innovation assessment. The SVM classifier's robust handling of high-dimensional feature spaces and effective decision boundary optimization proves particularly valuable in design evaluation scenarios where class boundaries may be subtle and complex. This combination leverages the representational power of deep learning while maintaining the interpretability and robustness advantages of traditional machine learning approaches. Our systematic evaluation of explainable AI techniques provides the first comprehensive comparison of XAI methods specifically for design evaluation applications. The superior performance of Layer-wise Relevance Propagation (95.6% localization accuracy, 93.4% expert acceptance) over Grad-CAM and Integrated Gradients establishes LRP as the preferred explainability technique for design assessment contexts. This finding has important implications for the broader application of XAI in creative domains, where detailed, hierarchical explanations align better with professional analytical thinking processes than simpler attention-based visualizations. The multi-dimensional evaluation framework developed in this study advances theoretical understanding of design innovation by providing a computational model that captures the complex, interrelated factors that contribute to creative excellence. The strong correlations observed between originality and overall innovation scores (r = 0.84) validate theoretical models of creativity that emphasize novelty as a fundamental component of innovation, while the significant contributions of aesthetic quality (r = 0.78) and functional effectiveness (r = 0.72) confirm the importance of holistic evaluation approaches that consider multiple quality dimensions simultaneously.

6.1 Practical Applications and Industry Impact

The practical implications of this research extend across multiple domains within the creative industries, offering transformative potential for design education, professional practice, and innovation management. In educational contexts, the system's ability to provide consistent, detailed feedback on student work addresses a critical need for scalable, objective assessment in design programs where traditional evaluation methods are often constrained by faculty time limitations and subjective variability. The 90% reduction in evaluation time achieved by our AI-assisted approach enables comprehensive assessment of large student portfolios that would be prohibitively time-consuming using traditional methods. More importantly, the interpretable explanations provided by the XAI components offer students actionable insights for improvement that go beyond simple quality scores. The 28% faster improvement in design quality observed in educational validation studies demonstrates that AI-assisted evaluation can accelerate learning and skill development in ways that traditional feedback methods cannot match. For professional design practice, the system offers significant value in multiple application scenarios. Design agencies can employ the framework for rapid initial assessment of creative concepts, enabling more efficient allocation of expert review time to the most promising ideas. The system's consistency and objectivity make it particularly valuable for design competitions and awards programs, where fair and reliable evaluation is essential for maintaining credibility and participant satisfaction[33]. The framework's cross-domain generalization capabilities (89-94% accuracy when applied across different design domains) suggest broad applicability across diverse creative contexts. This flexibility enables organizations to deploy a single evaluation system across multiple design disciplines, reducing training requirements and maintenance overhead while ensuring consistent quality standards across different creative teams and projects. Corporate innovation management represents another significant application domain, where the system can support systematic evaluation of design proposals, patent applications, and product development concepts. The objective, quantifiable assessment provided by the framework enables data-driven decision-making in innovation investment and resource allocation, potentially improving the efficiency and effectiveness of corporate R&D processes.

6.2 Limitations and Methodological Considerations

Despite the significant advances demonstrated in this study, several limitations must be acknowledged that constrain the generalizability and applicability of our findings. The dataset, while comprehensive within its scope, reflects primarily Western design traditions and aesthetic preferences, potentially limiting the framework's effectiveness in evaluating designs from different cultural contexts or aesthetic traditions. Cross- cultural validation studies would be necessary to establish the framework's applicability in global design contexts where different aesthetic principles and innovation criteria may apply. The expert annotation process, while achieving high inter-rater reliability (K = 0.87), remains fundamentally subjective and may reflect the particular perspectives and biases of the expert panel employed in this study. The 15 design professionals involved in annotation, despite their extensive experience, represent a limited sample of the broader design community and may not capture the full diversity of professional perspectives on design innovation and quality. The temporal stability of the evaluation framework presents another important consideration. Design trends, aesthetic preferences, and innovation criteria evolve continuously, potentially requiring periodic retraining or recalibration of the system to maintain relevance and accuracy. The framework's ability to adapt to changing design contexts and emerging aesthetic movements remains to be established through longitudinal studies. Technical limitations include the system's current focus on visual design evaluation, which may not fully capture non-visual aspects of design innovation such as user experience, emotional impact, or cultural significance. While our multi-dimensional evaluation framework addresses some of these concerns through functional effectiveness and market relevance assessments, more comprehensive evaluation approaches might require integration of additional data sources and evaluation modalities. The computational requirements of the framework, while reasonable for research applications, may present barriers to widespread adoption in resource-constrained environments. The DenseNet201 architecture requires significant GPU memory and

processing power, potentially limiting accessibility for smaller design organizations or educational institutions with limited computational resources.

6.3 Future Research Directions and Technological Evolution

The findings of this study open several promising avenues for future research that could further advance the state-of-the-art in computational design evaluation. Multi- modal evaluation approaches that integrate visual analysis with textual descriptions, user feedback, and contextual information represent a particularly promising direction for enhancing evaluation comprehensiveness and accuracy. The development of adaptive evaluation frameworks that can automatically adjust assessment criteria based on design domain, cultural context, or temporal trends would address current limitations related to generalizability and temporal stability. Machine learning approaches for meta-learning evaluation criteria could enable systems that continuously improve their assessment capabilities through exposure to new design examples and expert feedback. Real-time evaluation capabilities represent another important research direction, enabling integration of design assessment into interactive design tools and creative workflows. Such capabilities would support iterative design processes by providing immediate feedback on design modifications and enabling rapid exploration of creative alternatives. The extension of explainable AI techniques to provide not only explanations of current assessments but also generative suggestions for design improvement represents a significant opportunity for advancing AI-assisted creativity. A system that can identify specific design elements that need to be modified and propose improvement strategies will provide greater value for design professionals and students. Collaborative evaluation frameworks that combine AI assessment with human expertise in structured ways could leverage the complementary strengths of computational and human evaluation approaches. Such hybrid systems might achieve even higher accuracy and acceptance than purely automated approaches while maintaining the efficiency advantages of AI-assisted evaluation. The development of domain-specific evaluation models optimized for particular design disciplines could improve accuracy and relevance compared to general- purpose frameworks. Specialized models for product design, architectural design, or user interface design could incorporate domain-specific knowledge and evaluation criteria that enhance assessment quality within particular creative contexts.

6.4 Broader Implications for AI and Creativity Research

This research contributes to broader understanding of the relationship between artificial intelligence and human creativity, demonstrating that AI systems can effectively evaluate and interpret creative works when designed with appropriate attention to domain-specific requirements and interpretability needs. The success of our explainable AI approach suggests that transparency and interpretability are not merely desirable features but essential requirements for AI systems operating in creative domains where human understanding and acceptance are critical . The multi-dimensional evaluation framework developed in this study provides a computational model of design innovation that could inform broader research on creativity assessment and creative process understanding. The quantitative relationships identified between different quality dimensions offer insights into the structure of creative excellence that could guide both AI system development and human creativity research. The demonstrated effectiveness of hybrid AI architectures in creative evaluation contexts suggests that the future of AI-assisted creativity may lie not in replacing human judgment but in augmenting and enhancing human creative capabilities through transparent, interpretable computational tools. This perspective aligns with emerging paradigms of human-AI collaboration that emphasize complementary strengths rather than competitive replacement. The educational applications demonstrated in this study highlight the potential for AI systems to democratize access to high-quality creative education by providing consistent, detailed feedback that supplements traditional instruction methods. This capability could be particularly valuable in addressing educational inequalities and expanding access to quality design education in underserved communities. The framework's success in achieving both high accuracy and high interpretability challenges common assumptions about trade-offs between

AI system performance and explainability. Our results suggest that carefully designed AI systems can achieve superior performance precisely because they incorporate interpretability considerations from the outset, rather than treating explainability as a post-hoc addition to opaque models

7 Conclusion

This research presents a comprehensive explainable AI framework for automated design innovation assessment that successfully addresses critical challenges in computational creativity evaluation while establishing new standards for transparency and interpretability in AI-assisted design analysis. Through the systematic integration of advanced deep learning architectures, robust machine learning classification, and multiple explainable AI techniques, we have developed a system that achieves exceptional performance (97.8% accuracy) while providing interpretable explanations that meet the practical needs of design professionals and educators. The hybrid DenseNet201-SVM architecture demonstrates that combining the representational power of deep learning with the robustness and interpretability of traditional machine learning can yield superior results compared to end-to-end deep learning approaches in design evaluation contexts. This finding has important implications for AI system design in creative domains, suggesting that hybrid approaches may be more effective than purely deep learning solutions when interpretability and professional acceptance are critical requirements. Our systematic evaluation of explainable AI techniques establishes Layer-wise Relevance Propagation as the most effective approach for design evaluation applications, achieving 95.6% localization accuracy and 93.4% expert acceptance. This finding provides crucial guidance for implementing interpretable AI systems in creative contexts and demonstrates that detailed, hierarchical explanations align better with professional analytical thinking processes than simpler attention-based visualizations. The multi-dimensional evaluation framework developed in this study advances theoretical understanding of design innovation by providing a computational model that captures the complex, interrelated factors contributing to creative excellence. The strong correlations identified between different quality dimensions validate holistic approaches to design assessment while providing quantitative insights into the structure of creative evaluation that can inform both AI system development and design education practices. Practical validation through comprehensive user studies demonstrates significant improvements in evaluation efficiency (90% time reduction), consistency (37.6% improvement in inter-rater agreement), and educational effectiveness (28% faster student improvement) compared to traditional evaluation methods. These results establish clear evidence of practical utility and user acceptance that supports deployment in professional design contexts, educational institutions, and innovation management applications. The framework's cross-domain generalization capabilities and consistent performance across different design disciplines demonstrate broad applicability that extends beyond specific creative domains. This flexibility enables organizations to deploy unified evaluation systems across diverse design contexts while maintaining consistent quality standards and reducing training and maintenance requirements. Looking forward, this research opens several promising directions for advancing AI- assisted creativity and design evaluation. Multi-modal evaluation approaches that integrate visual analysis with textual descriptions and contextual information could further enhance assessment comprehensiveness. Adaptive evaluation frameworks that automatically adjust assessment criteria based on cultural context and temporal trends could address current limitations related to generalizability. Real-time evaluation capabilities could enable integration into interactive design tools, supporting iterative creative processes through immediate feedback and guidance. The broader implications of this work extend beyond design evaluation to fundamental questions about the relationship between artificial intelligence and human creativity. Our results demonstrate that AI systems can effectively evaluate and interpret creative works when designed with appropriate attention to domain-specific requirements and interpretability needs. This suggests that the future of AI-assisted creativity lies not in replacing human judgment but in augmenting and enhancing human creative capabilities through transparent, interpretable computational tools. The educational applications demonstrated in this study highlight the democratizing potential of AI-assisted design evaluation, providing access to consistent, high-quality feedback that can supplement traditional instruction methods and address educational inequalities. The framework's success in achieving both high accuracy and high interpretability challenges common assumptions about trade-offs between AI system performance and explainability, suggesting that carefully designed systems can achieve superior performance precisely because they incorporate interpretability considerations from the outset. In conclusion, this research establishes a new paradigm for AI-assisted design evaluation that combines computational precision with human-interpretable insights, offering substantial potential for transforming design education, creative industry workflows, and innovation management practices. The framework provides a foundation for future research in computational creativity while delivering immediate practical value for design professionals, educators, and organizations seeking to enhance their creative evaluation capabilities through transparent, reliable AI assistance.

DECLARATIONS

Ethics approval and consent to participate

Not applicable.

Conflict of interest

No potential conflict of interest was reported by the authors.

Dataset to be available

All data generated or analysed during this study are included in this published article.

Consent for publication

Not applicable.

Funding

This work was supported by Zhejiang Province's ”14th Five-Year Plan” Teaching Reform Project (jg20240128); Ningbo Municipal ”161” Project for the Construction of Disciplines and Specialties in Higher Education Institutions (Ningbo Education [2025] No. 2),and Research Project of Zhejiang Commercial Education Research Association (2025ZSJY08).

Acknowledge

We sincerely thank your organization for its strong support for this research! The smooth progress of this research would not have been possible without the support and assistance of the Zhejiang Leading Geese Research and Development Program (Project No.:2023C01216), the National Natural Science Foundation of China (Project No. : 52475287), and the Engineering Research Center for Computer-Aided Product Innovation Design of the Ministry of Education.

Authors' information

Xianli Shen performed Supervision, Idea, Writing, and Proof-Reading. Xianli Shen performed Writing, formal Analysis, and Methodology. Lingyan Zhang performed Writing, methodology, implementation, and visualization. Muling Huang performed Writing, Model Creation, methodology, and Validation. Dongjun Wu performed Writing, Formal Analysis, and Validation. Muling Huang performed Writing, Data curation, and interpretation. Lingyan Zhang Performed writing, software, visualization, and writing, data curation, and Funding.

References

[1] Goldschmidt, G., Tatsa, D.: How good are good ideas? correlates of design creativity. Design Studies **26**(6), 593–611 (2005). https://doi.org/10.1016/j.destud.2005.02.004

[2] Sweeting, B., Sutherland, S.: Design's secret partner in research: Cybernetic practices for design research pedagogy. Systems Research and Behavioral Science **40**(5), 765–771 (2023)

https://onlinelibrary.wiley.com/doi/pdf/10.1002/sres.2974. https://doi.org/10.1002/sres.2974

[3] Markus, A.F., Kors, J.A., Rijnbeek, P.R.: The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics **113**, 103655 (2021). https://doi.org/10.1016/j.jbi.2020.103655

[4] Daryanavard Chounchenani, M., Shahbahrami, A., Hassanpour, R., Gaydadjiev, G.: Deep learning based image aesthetic quality assessment-a review. ACM Comput. Surv. **57**(7) (2025). https://doi.org/10.1145/3716820

[5] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012

[6] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007

[7] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.-Z.: Xai—explainable artificial intelligence. Science Robotics **4**(37), 7120 (2019). https://doi.org/10.1126/scirobotics.aay7120

[8] Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052

[9] Vahdati, S., Fathalla, S., Lange, C., Behrend, A., Say, A., Say, Z., Auer, S.: A comprehensive quality assessment framework for scientific events. Scientometrics **126**(1), 641–682 (2021). https://doi.org/10.1007/s11192-020-03758-1

[10] Shirke, S.A., Jayakumar, N., Patil, S.: Design and performance analysis of modern computational storage devices: A systematic review. Expert Systems with Applications **250**, 123570 (2024). https://doi.org/10.1016/j.eswa.2024.123570

[11] Malik, A., Vaidya, G., Jagota, V., Eswaran, S., Sirohi, A., Batra, I., Rakhra, M., Asenso, E.: Design and evaluation of a hybrid technique for detecting sunflower leaf disease using deep learning approach. Journal of Food Quality **2022**(1), 9211700 (2022)

https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/9211700. https://doi.org/10.1155/2022/9211700

[12] Oxman, R.: Thinking difference: Theories and models of parametric design thinking. Design Studies **52**, 4–39 (2017). https://doi.org/10.1016/j.destud.2017.06.001. Parametric Design Thinking

[13] Gopsill, J., Giunta, L., Goudswaard, M., Snider, C., Hicks, B.: Data mining prototyping knowledge graphs for design process insights. Journal of Engineering Design **0**(0), 1–27 (2024) https://doi.org/10.1080/09544828.2024.2302746. https://doi.org/10.1080/09544828.2024.2302746

[14] Josifidis, K., Supic, N.: (are) institutions more important than innovation? Journal of Economic Issues **55**(2), 334–341 (2021) https://doi.org/10.1080/00213624.2021.1908086. https://doi.org/10.1080/00213624.2021.1908086

[15] Hay, L., Duffy, A.H.B., McTeague, C., Pidgeon, L.M., Vuletic, T., Grealy, M.: A systematic review of protocol studies on conceptual design cognition: Design as search and exploration. Design Science **3**, 10 (2017). https://doi.org/10.1017/dsj.2017.11

[16] Cash, P., Kreye, M.: Exploring uncertainty perception as a driver of design activity. Design Studies **54**, 50–79 (2018). https://doi.org/10.1016/j.destud.2017.10.004

[17] Yilmaz, S., Seifert, C.M.: Creativity through design heuristics: A case study of expert product design. Design Studies **32**(4), 384–415 (2011). https://doi.org/10.1016/j.destud.2011.01.003

[18] Krish, S.: A practical generative design method. Computer-Aided Design **43**(1), 88–100 (2011). https://doi.org/10.1016/j.cad.2010.09.009

[19] Avsec, S., Jagiełło-Kowalczyk, M., Żabicka, A.: Enhancing transformative learning and innovation skills using remote learning for sustainable architecture design. Sustainability **14**(7) (2022). https://doi.org/10.3390/su14073928

[20] Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis (2018)

[21] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

[22] Petch, J., Di, S., Nelson, W.: Opening the black box: The promise and limitations of explainable machine learning in cardiology. Canadian Journal of Cardiology **38**(2), 204–213 (2022). https://doi.org/10.1016/j.cjca.2021.09.004. Focus Issue: New Digital Technologies in Cardiology

[23] Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017)

[24] Holzinger, A., et al.: What do we need to build explainable AI systems for the medical domain? (2017)

[25] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps (2013)

[26] Ying, W., Zhang, L., Luo, S., Yao, C., Ying, F.: Simulation of computer image recognition technology based on image feature extraction. Soft Computing **27**(14), 10167–10176 (2023). https://doi.org/10.1007/s00500-023-08246-1

[27] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE **10**(7), 1–46 (2015). https://doi.org/10.1371/journal.pone.0130140

[28] Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.-R.: Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition **65**, 211–222 (2017). https://doi.org/10.1016/j.patcog.2016.11.008

[29] Arras, L., Horn, F., Montavon, G., Müller, K.-R., Samek, W.: ”what is relevant in a text document?”: An interpretable machine learning approach. PLOS ONE **12**(8), 1–23 (2017). https://doi.org/10.1371/journal.pone.0181142

[30] Ying, W., Zhang, L., Luo, S., Yao, C., Ying, F.: Simulation of computer image recognition technology based on image feature extraction. Soft Computing **27**(14), 10167–10176 (2023). https://doi.org/10.1007/s00500-023-08246-1

[31] Huang, L., Yao, C., Zhang, L., Luo, S., Ying, F., Ying, W.: Enhancing computer image recognition with improved image algorithms. Scientific Reports **14**, 13709 (2024). https://doi.org/10.1038/s41598-024-64193-3

[32] Merdjanovska, E., Rashkovska, A.: Comprehensive survey of computational ecg analysis: Databases, methods and applications. Expert Systems with Applications **203**, 117206 (2022). https://doi.org/10.1016/j.eswa.

2022.117206

[33] Tovey, M., Porter, S., Newman, R.: Sketching, concept development and automotive design. Design Studies **24**(2), 135–153 (2003). https://doi.org/10.1016/S0142-694X(02)00035-2