International Journal of Innovative Technology and Interdisciplinary Sciences
https://journals.tultech.eu/index.php/ijitis
ISSN: 2613-7305
Volume 8, Issue 2
DOI: https://doi.org/10.15157/IJITIS.2025.8.2.355-373
Received: 11.05.2025; Revised: 06.07.2025; Accepted: 19.07.2025

Research Article

Assessing Creativity in Text-to-Image Generation: A Quantitative Analysis using Structured Human Rating Metrics

Ramya Mandava*

College of Computing, Georgia Institute of Technology, Atlanta, United States of America
***ramyamresearcher@gmail.com**

Abstract

This research examines the creativity of text-to-image (T2I) generation models using a systematic human rating framework to evaluate four important dimensions of creativity: originality, relevance, aesthetic appeal, and imaginativeness. The advanced development of generative AI tools DALL•E 2, Mid journey, and Stable Diffusion creates subjective barriers to measuring their creative output. This evaluation analyses 100 pictures generated by DALL•E 2 versus Midjourney versus Stable Diffusion through testing a wide spectrum of commands from various artistic domains. The evaluation demonstrates that Mid journey offers better artistic results than DALL•E 2 and Stable Diffusion when comparing artistic achievements between the models. DALL•E 2 stands out for its relevance because it produces prompts with extremely strong semantic alignment to the provided instructions. The total creativity score for Stable Diffusion falls below its rivals but the model presents occasional improvements in originality. The framework quality shows itself through high agreement among evaluators. The evaluation needs multiple assessment methods to identify distinctive creative abilities of T2I models while providing important guidance for AI development in creative domains in the future. The research has established comprehensive evaluation standards that future investigations in creative AI must follow because of the essential need for methodological rigor.

Keywords: Text-to-Image Generation; Creativity Assessment; Human Evaluation; Midjourney; DALL•E 2; Stable Diffusion; Aesthetic Computing; Generative AI; Multimodal Models; Computational Creativity.

INTRODUCTION

The emerging trends of artificial intelligence (AI) have produced T2I generation as a game-changing innovation [1]. The image generation capabilities of DALL•E 2 along with Midjourney and Stable Diffusion produce various creative visual outputs from written input []2. The union of words with visual creativity provides fresh development prospects which benefit advertising entertainment and art and design among other sectors [3]. The current standard evaluation methods show limited success in measuring AI-produced image creativity due to their inability to assess creativity attributes like originality, pertinence, beauty, and inventiveness [2, 4].

https://doi.org/10.15157/IJITIS.2025.8.2.355-373

The evaluation of AI-generated content exceeds mere technical performance and follows the initial input instructions [4]. Users evaluate creative images by examining their uniqueness as well as their alignment with phrasing and their visual appeal and their abstract or innovative characteristics [3]. Existing evaluation techniques, which are based mostly on quantitative scores such as FID (Fréchet Inception Distance) or IS (Inception Score), tend to fall short in capturing the multi-faceted aspects of creativity [5]. These scores tend to centre around image quality and do not necessarily account for the more qualitative, subjective aspects that are part of creative output [6].

The impetus for this work is to fill the gap by proposing a more holistic, human-oriented evaluation framework to measure the creativity of T2I models [7]. Through the use of structured human rating measures, the study attempts to analyse how various models score along primary creative dimensions: originality, relevance, aesthetic quality, and imaginativeness [7]. Through a quantitative evaluation, this research not only assesses the creative products of AI systems but also seeks to aid in the creation of more sound, reliable instruments for measuring AI creativity in future endeavours [8].

This research work makes an exclusive contribution by providing a close, data-driven methodology for assessing the creativity of AI-generated images that includes inter-rater reliability and internal consistency to authenticate the human ratings [8-10]. The results of this study will offer significant knowledge on the merits and demerits of various T2I models that will help shape the future of AI development as well as its inclusion in creative processes. Additionally, it will influence the debate about AI-generated material, beyond purely technical evaluation towards a more integrated realization of the role of AI within the creative process [9, 11].

Background

Text-to-image (T2I) synthesis has made tremendous progress over the past few years, with systems like DALL•E, Midjourney, and Stable Diffusion making it possible to automatically generate images of good quality from text prompts [9]. These generation models utilize advanced algorithms, especially deep learning methods like transformers and diffusion models, to fill the semantic gap between text and visual representation [10]. The creativity and quality of produced images, however, differ between models and are commonly assessed subjectively, resulting in inconsistent judgments. With increasingly widespread adoption of AI-generated imagery across industries like digital art, design, advertising, and entertainment, it is ever more important to create standardized and trustworthy methodologies for assessing creativity [11-13].

Motivation

Present assessment tools for AI content mostly depend on automatic systems or human evaluation. Automatic assessment tools like Inception Score and Fréchet Inception Distance (FID) offer qualitative metrics about image quality but fail to detect crucial creative features such as level of originality alongside relevance and aesthetic value and imaginative quality [11]. The understanding of text-to-image model creativity depends on these core dimensions which also determine their industrial applications. Public and creative practitioners need standardized methods to assess creativity within these processes as they gain increasing exposure to these models. Human evaluation using creativity dimensions gives a qualitative means to understand the creative output of AI systems [13].

Special Contribution

The research brings a systematic quantitative technique to computational creativity through its text-to-image synthesis assessment method [14]. This research implements human raters to evaluate the creative outcomes of leading text-to-image models (DALL•E 2, Midjourney, and Stable Diffusion) based on the evaluation factors of originality, relevance, aesthetic value and imaginativeness. An evaluation of human reliability and slight creative distinctions between models was achieved through ICC assessments of inter-rater agreement and Cronbach's α calculations of internal consistency. The research delivers vital knowledge regarding model strengths and barriers for future advancements in model development assessment practices [12, 15]. The research results demonstrate the need to use systematic multi-dimensional assessment methods for AI creativity measurement because they guide future investigations about generative AI applications in creative fields [13, 16].

Objectives of the Research

- A numerical system needs development for measuring the creative output of images generated by AI.
- The study compares creative abilities between DALL•E 2, Midjourney, and Stable Diffusion at the top of the text-to-image model’s market.
- A well-structured framework needs development which will enable people to evaluate creativity across defined parameters [17].

LITERATURE REVIEW

Authors at [18] presented a two-aspect evaluation framework specially created for human image synthesis with T2I models. Their framework partitioned assessment criteria into two primary dimensions: image-oriented properties like aesthetics and realism, and text-conditioned properties like concept coverage and fairness. For supporting automated evaluations, they proposed a new aesthetic score prediction model and made available a dataset with low-quality area annotations in generated human images [19]. Their work also stressed the importance of fairness assessments, pointing out gender, racial, and age biases in model outputs, thereby calling for more contextually and ethically conscious generative models [13].

In the [14] has been investigated the T2I generation's creative aspect, with special emphasis on the practice of prompt engineering and the wider social environment of AI art [20]. The research contended that the conventional product-based understanding of creativity was inadequate to grasp the dynamics of AI art, particularly where community involvement and prompt curating are essential. Applying Rhodes' four P creativity model (Person, Process, Press, Product), the writer underscored text-to-image creation's interpersonal and sociocultural aspects and challenged the evaluation of such creativity via current static measurements [14, 21].

Authors at [15] addressed the challenge of measuring generative image quality in terms of human preference modelling. They created "Image Reward"—the first large-scale, general-purpose reward model trained on 137,000 expert-rated image pairs [22]. Their work bested current evaluation metrics and made Reward Feedback Learning (ReFL), a new tuning procedure for human-guided improvement of diffusion models, possible. Their results highlighted the power of human-centered preference data for both model training and evaluative precision [15, 23].

Authors at [16] performed a comparative analysis between human-created and machine-created images with DALL•E using Consensual Assessment Technique (CAT) and the Turing Test (TT) as standards for assessing combinational creativity. They proposed two statistical evaluation measures: Coincident Rate (CR) and Average Rank Variation (ARV) to examine metric consistency [24]. Their research found that GIQA, along with three other metrics tested, attained the best correlation with human-based benchmarks, indicating its promise as a feasible measure for automated creativity evaluation, especially for engineering and design purposes [16, 25].

In the [17] were investigated the rising necessity of commitment quality parameters in T2I models, especially for scenarios where precise control over scene parameters (light, material, object positioning) was anticipated [26, 27]. They suggested a taxonomy that grouped T2I image quality measures into two wide categories: compositional quality and overall image quality. Their paper also spoke about benchmark datasets for text-to-image testing and drew attention to the incompatibility of old metrics of rendering such as SSIM or PSNR with the needs of generative, prompt-driven image synthesis [17, 26].

In the [18] has been analysed the application of text-to-image generation in creative writing processes. His research investigated the ways in which authors utilize AI-produced imagery for inciting narrative progress, aiding world-building, and augmenting visual imagination. The dissertation emphasized that T2I systems were used as inspiration resources as well as co-creative collaborators and assisted authors in bridging the chasm between abstract text idea and tangible visual execution. Ivan underlined that such cross-modal interaction had the power to re-design conventional writing processes, even while recognizing a demand for image fidelity refinement as well as better alignment with prompts [18].

Authors at [19] conducted an image perception study about T2I system products to analyse human-AI co-creation communication processes. Subjects in the study gave meanings while assigning emotions and intentional qualities to artificial work despite knowing it stemmed from artificial origins. The presence of social interpretation shows that people apply their own precepts to co-creative processes. The research established that meaningful human–AI interaction along with timely engineering and control mechanisms serve to prevent over-attribution of agency to machines [19].

Authors at [20] investigated T2I generation applications for architectural design processes starting from initial conceptualization stages. Architects have shown according to their study they can recreate forms and spatial concepts by using generative models that also reproduce stylistic elements from verbal instruction. According to the research T2I tools enabled rapid design testing together with conceptual development and explorative design functions mostly within the ideation and brainstorming phases. While the researchers documented their findings, they concluded that the generated outputs did not reach sufficient technical accuracy or structural feasibility levels for future design stages. Building designers need domain-adapted tools alongside architectural software integration to use these tools effectively for realistic applications [20, 28].

Research gap

Various research holes prevent T2I technology from achieving its full creative, evaluative and technical potential despite recent significant development in text-to-image generation technology. Although research like that of [13, 19] has introduced novel evaluation frameworks and taxonomies for image quality assessment, there remains no single, domain-independent metric capable of capturing both aesthetic and semantic fidelity simultaneously, particularly in subjective domains like creative writing or architecture design. Additionally, current evaluation models are inclined to observe end-product quality without satisfyingly factoring in the interactive and co-creative nature of human-AI collaboration, as noted by [14]. While preference-based models such as Image Reward [15, 16] have demonstrated potential, they continue to have generalizability and interpretability issues across creative domains. In addition, the studies expose a lack of attention to the ethical and social-cultural aspects of T2I outputs, including fairness and representation bias, which [13] merely mentioned but were not thoroughly examined in real-world applications such as architecture [18, 20]. There is also a significant lack of research that investigates how users with non-technical or novice backgrounds use and learn to work with T2I tools, yet this is most important for democratizing access to this technology. More generally, these shortcomings indicate the necessity of more interdisciplinary, user-focused, and ethically informed research to guarantee that T2I systems develop as genuinely inclusive, creative, and context-aware tools [29].

PROPOSED METHOD

In this research, a systematic methodology was used to quantitatively measure creativity in text-to-image generation models, see Figure 1. The methodology was developed to guarantee systematic data collection, regular human assessment, and statistical analysis based on dimensions related to creativity [30]. The main goal was to analyse the perceived creativity of images produced by AI models from text inputs, employing systematic human rating metrics.

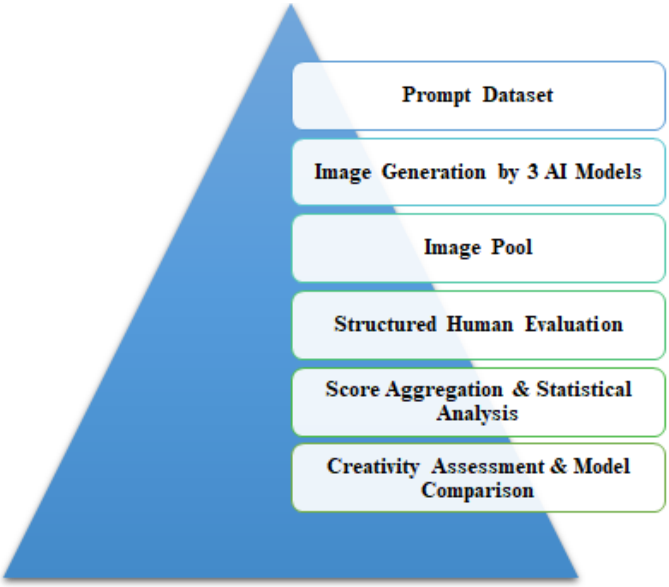
Prompt Dataset
Image Generation by 3 AI Models
Image Pool
Structured Human Evaluation
Score Aggregation & Statistical Analysis
Creativity Assessment & Model Comparison

Figure 1. Research Framework

Research Design

The research utilized a mixed-methods approach with a focus on quantitative assessment. A within-subject design was employed, where the multiple image outputs obtained based on the same prompt and created with various AI models were evaluated by the participants [31]. Using this approach enabled comparative analysis and reduced individual prompt interpretation bias.

Selection of Text Prompts

A handpicked dataset of 100 open-ended and diverse textual prompts was gathered from a pool of available creative writing datasets and manually designed descriptions. Prompts were chosen to cover all kinds of genres like surrealism, science fiction, nature, abstract, and fantasy. Each prompt was crafted to provide broad room for creative interpretation by the image generation models [32].

Text-to-Image Generation Models

Three cutting-edge text-to-image generation models were chosen for testing: DALL•E 2, Midjourney, and Stable Diffusion. One image was produced by each model for every prompt, yielding a total of 300 images (100 prompts × 3 models). All the images were created in controlled conditions to eliminate differences in resolution and style settings

Human Rating Procedure

Rater Recruitment and Training

Thirty human raters were enrolled through academic email lists and a verified crowdsourcing website. They went through a 90-minute training session of the four creativity dimensions (Originality, Relevance, Aesthetic Appeal, Imaginativeness) using definitions and examples taken from existing studies. The raters looked at "anchor-point" pictures depicting scores 1, 4, and 7 on the 7-point rating scale, heard about borderline examples, and were allowed to rate ten image-prompt pairs as practice. Trainers gave feedback until the scores of each rater were within one point of the gold-standard anchors, with consistent interpretation of the scale [33].

Blind, Independent Rating Procedure

To avoid bias, all images were shown without any mention of the generating model. Each rater was assigned 50 different image-prompt pairs randomly, and each pair was rated by at least five different raters. Raters worked independently and asynchronously, with no communication allowed during scoring. This design balanced workload, maximized coverage, and allowed strong inter-rater reliability analysis.

Rating Scale and Anchor-Point Definitions

For each creative component, raters assigned a score to each picture on a 7-point Likert scale, where 1 represents extremely low and 7 represents very high. To secure the scale:

- **Originality:** 1 = resembles generic stock imagery; 4 = some novel elements but partly derivative; 7 = entirely novel composition.
- **Relevance:** 1 = irrelevant to prompt; 4 = partially addresses prompt; 7 = fully captures and enhances prompt meaning.
- **Aesthetic Appeal:** 1 = visually jarring; 4 = adequate with minor flaws; 7 = professional, polished.
- **Imaginativeness:** 1 = strictly literal; 4 = some creative deviation; 7 = highly metaphorical or abstract.

These anchor-point examples provided objective benchmarks to minimize subjective drift over time.

Questionnaire and Rubric Structure

The web rubric showed one item per dimension with its definition and two exemplar thumbnails (low and high anchor). Raters clicked a radio button for their score and were able to go back to definitions anytime. By inserting both textual definitions and visual anchors directly into the questionnaire, we operationalized each creativity dimension clearly and minimized ambiguity in what, for example, "Originality = 5" means in practice.

Ensuring Objectivity and Reliability

To further guard against subjectivity, we employed:

- **Blind Rating:** The identity of the models are concealed throughout.

- **Mid-Study Calibration:** Every rater reassessed a mini-set of five photos halfway through; variations more than ±1 point led to a quick recalibration session utilising anchor images.
- **Inter-Rater Reliability Checks:** We calculated the ICC for each dimension, and an adequate agreement was defined as ICC > 0.75. Raters were classified as outliers and removed from the final analysis if their total ICC was less than 0.5.
- **Internal Consistency:** The four dimensions' Cronbach's alpha was computed, and $\alpha > 0.7$ indicated that the scale items accurately assessed creative components.

Blind and independent scoring, statistical reliability assessments (ICC and Cronbach's alpha), frequent recalibration, stringent training, and precise anchor-point definitions all helped to reduce human bias and guarantee that our ratings served as a strong basis for further quantitative analysis.

Reliability and Validity Measures

To allow for inter-rater reliability, ICC was calculated for each of the dimensions. Cronbach's alpha was also determined to gauge internal consistency of the rating scales. To establish validity of the human rating instrument, pilot study was done before the main data collection [34].

In order to confirm the reliability of human ratings, we estimated the Intraclass Correlation Coefficient (ICC), a statistical technique used to quantify the level of agreement or consistency between several raters. ICC has been calculated through equation (1) and is especially appropriate in this research since it compares inter-rater reliability by examining the variance among raters against the total variance of data.

$$ICC = \frac{\text{Between-group variance}}{\text{Total variance}} \tag{1}$$

Where:

- ***Between-group variance***, represents variation owing to differences between raters.
- ***Total variance***, consists of between-group variance and within-group variance (resulting from individual differences or error).

The ICC is a value between 0 and 1, where higher values mean greater agreement among raters. The higher the ICC value, the more consistent the ratings of different raters are; the lower the ICC value, the more variability there is in the ratings.

Additionally, to test the internal consistency of the rating scale, Cronbach's alpha was calculated. This coefficient assists in determining how well the individual items within the scale (e.g., Originality, Relevance, Aesthetic Appeal, and Imaginativeness) relate to one another, so that the scale is measuring a single construct. The Cronbach's alpha is calculated through equation (2).

$$\alpha = \frac{N \cdot \overline{C}}{\overline{V} + (N-1) \cdot \overline{C}} \tag{2}$$

Where:

- **N** is the number of items,
- $\mathbf{\bar{C}}$ is the average covariance between items,
- $\mathbf{\bar{V}}$ is the average variance of each item.

By applying both ICC and Cronbach's alpha, we are confident that the human raters' ratings are consistent (reliable) as well as internally consistent across the various dimensions of creativity.

Data Analysis

Descriptive statistics were calculated to consolidate the ratings per model by dimensions of creativity. One-way repeated measures ANOVA was conducted to identify statistically significant model-level differences. Post-hoc pairwise comparisons with Bonferroni adjustment were employed to identify which specific model-level differences exist [35, 36]. To assess relationships between different creativity dimensions correlation analysis was performed. Figure 2 depict the block diagram of statistical analysis.

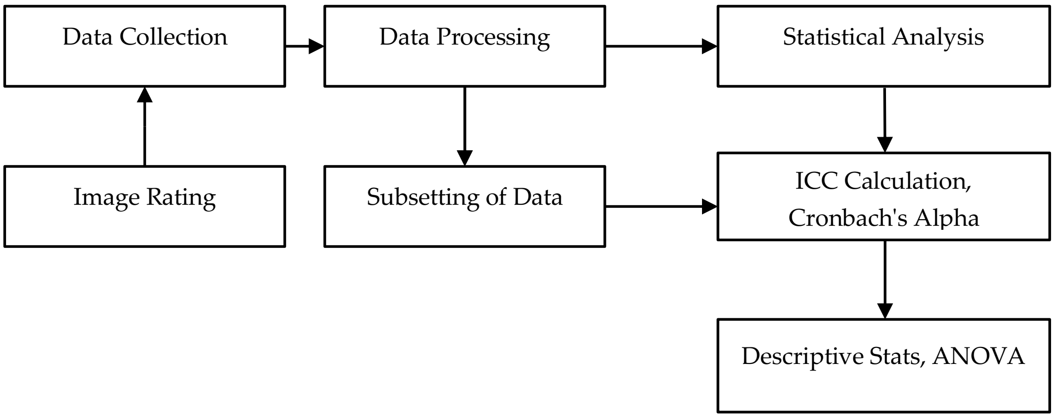

Figure 2. Block diagram of statistical analysis

In this research, ICC is applied instead of other methods like Pearson's Correlation and the Kappa Coefficient since it has the ability to measure the extent of agreement between multiple raters and therefore is best suited to measure inter-rater reliability where ratings are continuous (1-7 Likert scale). Pearson's Correlation measures the relationship between two variables without accounting for the reliability of multiple observers, while Kappa is designed for categorical data and does not deal with continuous variables well. The research design is in a coherent sequence: Data Collection involves generating images in reaction to prompts and having human raters assign ratings; Data Processing is to maintain consistency, compute ICC for inter-rater consistency, and calculate descriptive statistics (mean, standard deviation) for each creativity dimension; Statistical Analysis uses one-way repeated measures ANOVA to detect significant differences in creativity ratings, to which post-hoc tests (e.g., Bonferroni) are used to determine specific model differences; and Results Interpretation concludes on the basis of statistical significance, including ICC values and Cronbach's alpha. ICC is the most suitable for this research because it is able to efficiently handle continuous and categorical data with strong and consistent results.

Ethical Considerations

Every participant provided their informed consent to participate before joining the study. Researchers adhered to ethical standards of human research subject participation after obtaining authorization from their institutional review board at their affiliated academic institution [35].

RESULTS AND DISCUSSION

The results of human scores on text-to-image outputs from DALL•E 2, Midjourney, and Stable Diffusion appear in this section where an analysis is also provided. Assessments were conducted in four creative dimensions which consisted of originality, relevance, aesthetic appeal and imaginativeness. The research used descriptive statistics and various inter-rater reliability indices and inferential statistical tests to analyse results which distinguished model performances.

Descriptive Statistics of Creativity Ratings

Table 1 and Figure 2 reveals the averaged creativity scores obtained from user evaluations across all dimensions for Stable Diffusion, Midjourney and DALL•E 2. Midjourney led the rating metrics across all categories including aesthetic appeal (6.02) and imaginativeness (5.83) which indicates its capability to produce visually attractive creative images. DALL•E 2 exhibited strong Relevance (5.65) abilities because it could generate images that closely matched the instructions provided by users.

The ratings for Stable Diffusion remained at the low end throughout most categories with its lowest mark recorded in Aesthetic Appeal at 5.10 because these models failed to produce appealing images. Notwithstanding this, Stable Diffusion had equally balanced performance in all the dimensions, though it was not superior to the other models in any given category. Research confirms that Midjourney stands out in terms of creativity by delivering aesthetically pleasing and original results.

Table 1: Mean (M) of Creativity Scores by Model

Creativity Dimension	DALL·E 2	Midjourney	Stable Diffusion
Originality	5.21	5.47	5.12
Relevance	5.65	5.41	5.22
Aesthetic Appeal	5.38	6.02	5.10
Imaginativeness	5.14	5.83	5.05

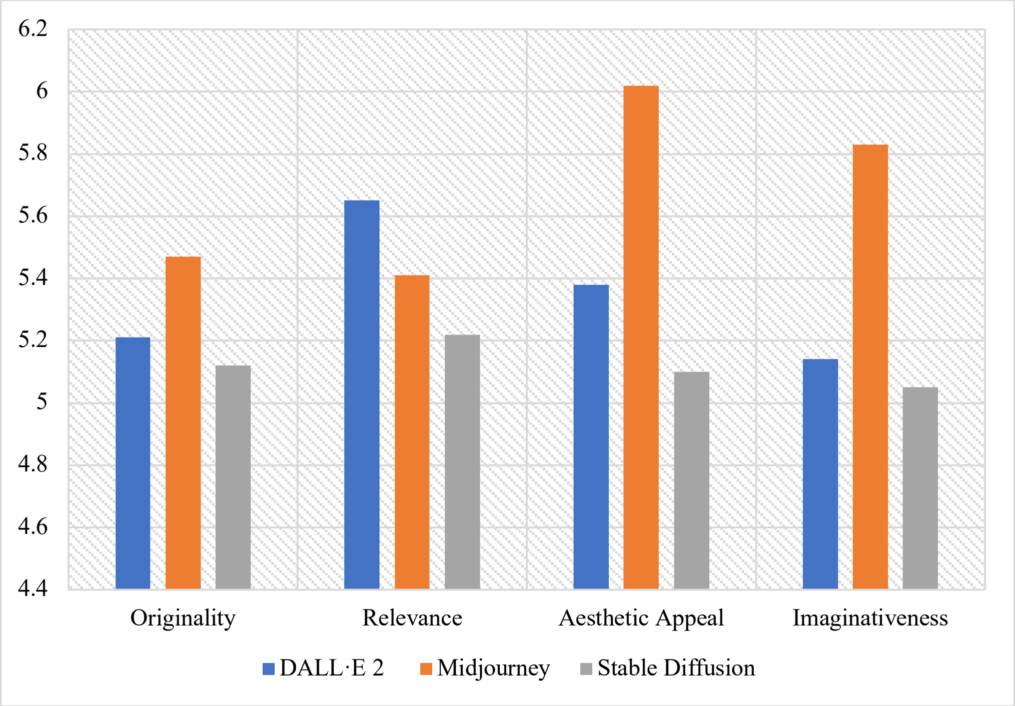

Figure 2. Mean (M) of creativity scores by model

Inter-Rater Reliability and Internal Consistency

The assessment ratings for creativity dimensions achieve highly reliable results because human raters show strong internal consistency (Cronbach's α) and consistent judgment across raters (ICC). The measure of agreement for the raters concerning image assessment has produced ICC values between 0.76 and 0.83 for all dimensions. Table 2 depict the results of ICC and Cronbach.

Table 2. Inter-Rater Reliability (ICC) and Internal Consistency (Cronbach's α)

Creativity Dimension	ICC (Average Measures)	Cronbach's α
Originality	0.78	0.84
Relevance	0.81	0.87
Aesthetic Appeal	0.83	0.88
Imaginativeness	0.76	0.82

Furthermore, Figure 3 depicts the graphical representation of Inter-Rater Reliability (ICC) of our research work.

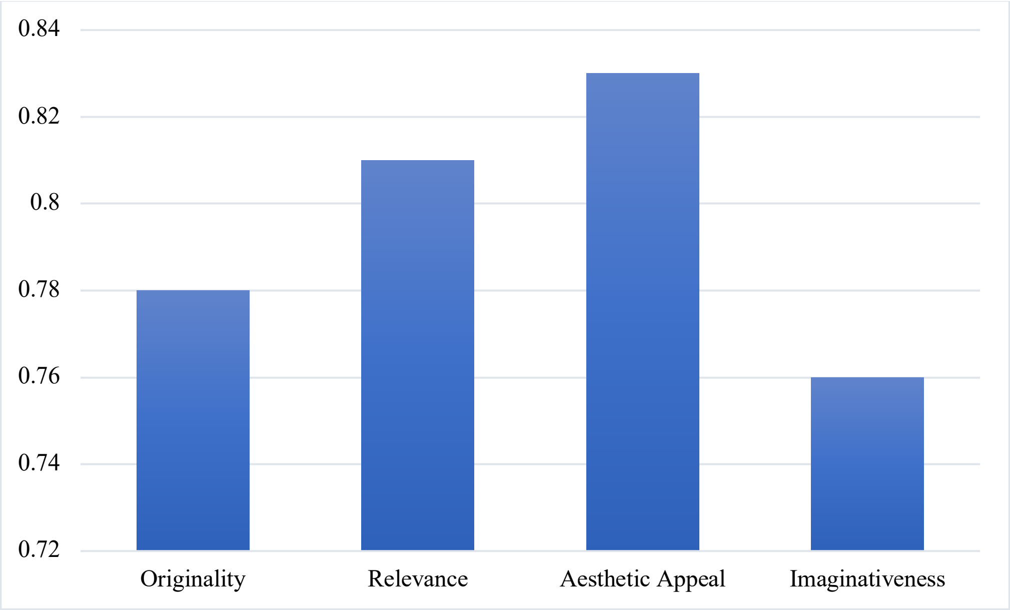

Figure 3. Graphical representation of Inter-Rater Reliability (ICC)

The Aesthetic Appeal dimension reaches the highest ICC score of 0.83 showing that evaluators deeply concurred on image visual quality. The four dimensions of creativity exhibit high internal consistency based on Cronbach's α that ranges from 0.82 to 0.88. The maximum Cronbach's α is also for Aesthetic Appeal (0.88), which implies that the items applied to measure this dimension are strongly consistent. In general, these findings indicate that the dimensions of creativity are accurately measured and that the assessment framework is successful in capturing human raters' subjective opinion.

Inferential Analysis: ANOVA and Post-hoc Tests

A one-way repeated measures ANOVA indicated statistically significant differences between the models in all creativity dimensions ($p <.01$). Post-hoc tests with Bonferroni correction indicated that:

- Midjourney far surpassed Stable Diffusion and DALL•E 2 on aesthetic quality and imaginativeness ($p < .001$).
- DALL•E 2 was much higher than Stable Diffusion in relevance ($p <.05$), but not otherwise.

These results imply that although all models can generate creative images, they vary according to the type of creativity expressed. Midjourney is more adept at expressing and imaginative visual composition, while DALL•E 2 is better suited to accurately reproducing the contents of the text prompt.

Correlation Between Creativity Dimensions

The correlation matrix for the four dimensions of creativity (Originality, Relevance, Aesthetic Appeal, and Imaginativeness) shows significant relationships between them, see

Table 3 and Figure 4. The highest correlation is found between Imaginativeness and Aesthetic Appeal (0.69), which indicates that images rated as more imaginative are also likely to be rated higher in aesthetic appeal.

Table 3. Correlation Matrix Between Creativity Dimensions

Dimension	Originality	Relevance	Aesthetic Appeal	Imaginativeness
Originality	1.00	0.42	0.53	0.61
Relevance	0.42	1.00	0.48	0.50
Aesthetic Appeal	0.53	0.48	1.00	0.69
Imaginativeness	0.61	0.50	0.69	1.00

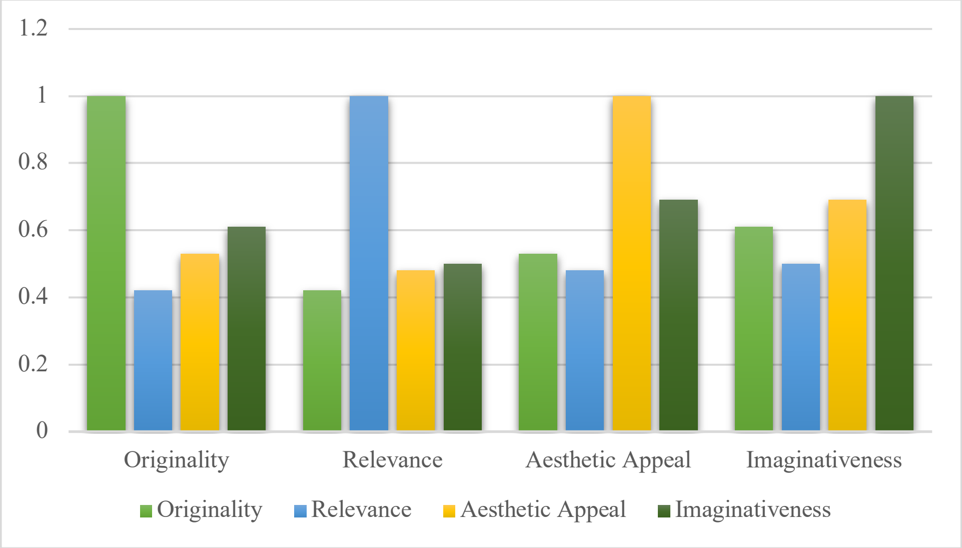

Figure 4. Correlation Matrix Between Creativity Dimensions

There is a moderate correlation of Originality and Imaginativeness (0.61), which suggests that images ranked as more original are also rated as more imaginative. But Originality has comparatively lower correlations with Relevance (0.42) and Aesthetic Appeal (0.53), suggesting that although originality matters, it does not highly correlate with how well the image fits the prompt or its attractiveness. Relevance also has moderate correlations with the other dimensions, indicating that relevance, although important, is not highly correlated with aesthetic or imaginative attributes. Overall, these correlations indicate that although each dimension is related, they each uniquely contribute to the overall measure of creativity in AI-generated images.

Table 4 and Figure 5 discloses significant genre-based differences in creativity scores between AI models, with different areas of strength. Midjourney excels consistently in creative styles such as Abstract, Sci-Fi, and Surrealism, with particular strengths in aesthetic beauty and imaginativeness, which is consistent with its excellent artistic realization capabilities and aptness for visually oriented tasks. DALL•E 2 demonstrates exceptional performance in the Nature category because it possesses the highest relevance score which indicates better prompt compatibility alongside its ability to generate realistic appropriate results. The Fantasy category indicates Stable Diffusion achieves balanced performance yet the model demonstrates no superior scores across any domain since it maintains a general approach instead of specialized functions. The study reveals that AI image creativity depends heavily upon design choices in models together with appropriate genre selection.

Table 4. Average creativity ratings by genre

Genre	Model	Originality	Relevance	Aesthetic Appeal	Imaginativeness
Abstract	Midjourney	5.52	5.34	6.20	6.10
Nature	DALL·E 2	5.10	5.85	5.38	5.20
Sci-Fi	Midjourney	5.65	5.25	6.00	5.95
Fantasy	Stable Diffusion	5.15	5.00	5.10	5.12
Surrealism	Midjourney	5.70	5.40	6.15	6.05

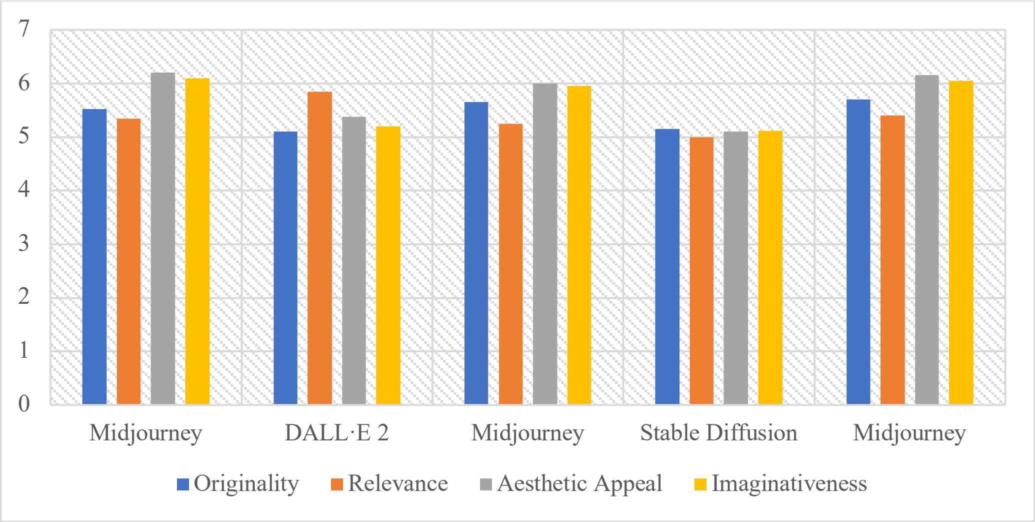

Figure 5. Average creativity ratings by genre

DISCUSSION

Systematic human judgments of AI-generated images deliver important insights about how various models are perceived by humans regarding their creative qualities. The artistic strengths and creative capabilities of Midjourney demonstrate its developed aesthetic skill sets most likely because of its training on artistic datasets and stylistic rendering methods.

The more significant relevance score of DALL•E 2 points to a better semantic match with prompts because the transformer-based architecture optimizes text-image consistency. Stable Diffusion showed versatility as a model but its overall strengths were moderate across every dimension which resulted in a balanced creative profile with few distinct features.

The findings from this research match recent academic findings indicating that artificial intelligence creativity exists in different forms based on algorithm specifications and learning datasets. Structured evaluation frameworks demonstrate high rating metric consistency in research on computational creativity according to the results of this study.

Research expansion should focus on two areas: examining creativity evaluation differences across cultures and improving experimental conditions through combination of text and emotional stimuli and expert artist assessment.

SUMMARY AND CONCLUSION

A large quantitative assessment of creativity emerged from systematic human ratings made across four core features: originality, relevance, beauty and imagineness when measuring text-to-image synthesis. Midjourney achieved superior results among the models that were evaluated alongside DALL•E 2 and Stable Diffusion by consistently demonstrating quality and imaginative capabilities when generating abstract and surrealistic and science fiction images. The system received the highest average assessments for originality along with artistic quality and imaginative ability indicating its production of visually striking conceptually advanced outputs. The semantic interpretation abilities of DALL•E 2 exceeded those of its competitors who were less capable of delivering relevance to generated images. The structured rating method exhibited stable reliability because of the high agreement levels shown by raters in all assessments. The scoring patterns showed that most AI-produced images continue to meet expectations by receiving positive assessments from human evaluators regarding creativity. Systematic human evaluation of AI-generated content proved practical through this research because it established essential markers for future innovative AI model and multimodal generative model development. Some limitations are as follows:

- Subjectivity in Human Ratings: Even bias in assessments with structured metrics and high inter-rater reliability, creativity is inherently subjective and may introduce.
- Limited Prompt Diversity: Even though the prompts show diversity they fail to represent complex scenarios specifically applicable to medical fields and technical and cultural artistic domains.
- Model Version Constraints: The research employed particular versions of DALL•E 2, Midjourney, and Stable Diffusion which could produce different performance results if updated versions were employed.
- Static Evaluation Approach: The research analysis focused exclusively on human judgments without incorporating automated creativity evaluation or explainable methods.

- Lack of Computational Metrics: Human evaluation served as the only assessment method while the study excluded both creativity measurement automation and explainability system integration.

Future research work will be focused on:

- Include Bigger and More Varied Prompt Sets: The addition of cross-cultural, technical and abstract prompts to the dataset would generate more extensive knowledge about model creativity.
- Blend Human and Automated Scoring: Mixing machine-based creativity and aesthetic scoring algorithms with human evaluation can create more thorough and scalable assessments.
- Longitudinal Studies of Creativity: Measuring model performance longitudinally and through updates may indicate trends in creative development or degradation.
- Investigate Domain-Specific Creativity: Research should evaluate creativity in various domains including education as well as advertising, entertainment and healthcare.
- User-Centred Evaluations: Researching how end-users (artists, designers, teachers) engage with and experience creativity in AI-generated content may provide useful insights for practical applications.
- Test Multimodal Outputs: Incorporating outputs with text, motion, or sound in addition to images would enable testing creativity in more sophisticated, multimedia scenarios.

CONFLICT OF INTERESTS

Author declares no conflict of interest

REFERENCES

1. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., ... & Ramanan, D. Evaluating text-to-visual generation with image-to-text generation. *Computer Vision – ECCV 2024: 18th European Conference*, Milan, Italy, September 29–October 4, **2024**. pp. 366-384.

2. Ko, H. K., Park, G., Jeon, H., Jo, J., Kim, J., & Seo, J. Large-scale text-to-image generation models for visual artists' creative works. 28th International Conference on Intelligent User Interfaces, IUI 2023 - Sydney, Australia, 27 Mar - 31 Mar **2023**. pp. 919-933.

3. Alhabeeb, S. K., & Al-Shargabi, A. A. Text-to-image synthesis with generative models: Methods, datasets, performance metrics, challenges, and future direction. *IEEE Access*, **2024**, 12, 24412-24427.

4. Turchi, T., Carta, S., Ambrosini, L., & Malizia, A. Human-AI co-creation: evaluating the impact of large-scale text-to-image generative models on the creative process. *End-User Development: 9th International Symposium, IS-EUD 2023*, Cagliari, Italy, June 6–8, **2023**, pp. 35-51.

5. Wu, H., Wu, X., Li, C., Zhang, Z., Chen, C., Liu, X., ... & Lin, W. T2i-scorer: Quantitative evaluation on text-to-image generation via fine-tuned large multi-modal models. *In Proceedings of the 32nd ACM International Conference on Multimedia*. China, October, **2024**, pp. 3676-3685.
6. Wiles, O., Zhang, C., Albuquerque, I., Kajić, I., Wang, S., Bugliarello, E., ... & Nematzadeh, A. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings. **2024**, arXiv preprint arXiv:2404.16820.
7. Lee, T., Yasunaga, M., Meng, C., Mai, Y., Park, J. S., Gupta, A., ... & Liang, P. S. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, **2023**, 36, 69981-70011.
8. Wan, Y., Xiao, L., Wu, X., Yang, J., & He, L. Imaginique Expressions: Tailoring Personalized Short-Text-to-Image Generation Through Aesthetic Assessment and Human Insights. Symmetry, **2024**, 16(12), 1608.
9. Bakr, E.M., Sun, P., Shen, X., Khan, F.F., Li, L.E., & Elhoseiny, M. HRS-Bench: Holistic, Reliable and Scalable Benchmark for Text-to-Image Models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, **2023**, pp. 19984-19996.
10. Aziz, M., Rehman, U., Safi, S.A., & Abbasi, A.Z. Visual Verity in AI-Generated Imagery: Computational Metrics and Human-Centric Analysis. **2024**, arXiv preprint arXiv:2408.12762.
11. Combs, K., Bihl, T. J., Gadre, A., & Christopherson, I. A human-factors approach for evaluating ai-generated images. *In Proceedings of the 2024 Computers and People Research Conference*, **2024**, pp. 1-9.
12. Aghazadeh, A., &Kovashka, A. CAP: Evaluation of Persuasive and Creative Image Generation. **2024**, arXiv preprint arXiv:2412.10426.
13. Chen, M., Liu, Y., Yi, J., Xu, C., Lai, Q., Wang, H., ... & Xu, Q. Evaluating text-to-image generative models: An empirical study on human image synthesis. **2024**, arXiv preprint arXiv:2403.05125.
14. Oppenlaender, J. The creativity of text-to-image generation. *In Proceedings of the 25th International Academic Mindtrek Conference*, Tampere, Finland, November 16–18, 2022, **2022**, pp. 192-202.
15. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., ... & Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, **2023**, 36, 15903-15935.
16. Wang, B., Zhu, Y., Chen, L., Liu, J., Sun, L., & Childs, P. A study of the evaluation metrics for generative images containing combinational creativity. *AI EDAM*, **2023**, 37, e11.
17. Hartwig, S., Engel, D., Sick, L., Kniesel, H., Payer, T., Poonam, P., ... &Ropinski, T.A Survey on Quality Metrics for Text-to-Image Generation. **2024**, arXiv preprint arXiv:2403.11821.
18. Ivan, A. S. G. (2024). Exploring the use of text-based image generation for creative writing. Doctoral dissertation, Nara Institute of Science and Technology.
19. Lyu, Y., Wang, X., Lin, R., & Wu, J. Communication in human–AI co-creation: Perceptual analysis of paintings generated by text-to-image system. *Applied Sciences*, **2022**, 12(22), 11312.
20. Albaghajati, Z.M., Bettaieb, D.M., & Malek, R.B. Exploring text-to-image application in architectural design: Insights and implications. *Architecture, Structures and Construction*, **2023**, 3(4), 475-497.

21. Dedeepya, P., Sowmya, P., Saketh, T.D., Sruthi, P., Abhijit P. and Praveen, S.P. Detecting Cyber Bullying on Twitter using Support Vector Machine," *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, Coimbatore, India, **2023**, pp. 817-822,

22. Swamy, B.N., Dedeepya, P., Sekhar, J.C., Pratap, V.K., Ananya K. and Sindhura, S. Brain Tumor Detection using RCNN and MobileNet, *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Coimbatore, India, **2023**, pp. 657-662,

23. Dedeepya, P., Karishma, D., Manuri, S.G., Raghuvaran, T. Shariff V. and Sindhura, S. Enhancing Cyber Bullying Detection Using Convolutional Neural Network, *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India, **2023**, pp. 1260-1267.

24. Dedeepya, P., Chiranjeevi, P., Narasimha, V., Shariff, V., Ranjith J. and Ramesh, J.V.N. Image Recognition and Similarity Retrieval with Convolutional Neural Networks, *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, Pudukkottai, India, **2023**, pp. 709-716.

25. Manasa, Y., Dedeepya, P., Sai, C.M., Navya, D., Timmasarti H.P. and Gangavarapu, M. Cyberbullying Tweets Detection Within Twitter Using CNN, *2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*, Jabalpur, India, **2024**, pp. 782-788.

26. Dedeepya, P., Yarrarapu, M., Kumar, P.P., Kaushik, S.K., Raghavendra P.N. and Chandu, P. Fake News Detection on social media Through a Hybrid SVM-KNN Approach Leveraging Social Capital Variables, *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, Salem, India, **2024**, pp. 1168-1175,

27. PBV, R.R., Koti Mani Kumar, T.N.S., Praveen, S.P., Sindhura, S., Al-Dmour N.A. and Islam, S. Optimizing Lung Cancer Detection: The Synergy of Support Vector Machine and Random Forest," *2024 International Conference on Decision Aid Sciences and Applications (DASA)*, Manama, Bahrain, **2024**, pp. 1-7,

28. Ponnaganti, N.D., Kumar, T. N. S. K. M., Praveen, S.P., Sindhura, S., Al-Dmour N.A. and Islam, S. A Robust SVM Framework for Heart Disease Detection Utilizing Advanced Feature Selection Techniques, *2024 International Conference on Decision Aid Sciences and Applications (DASA)*, Manama, Bahrain, **2024**, pp. 1-7,

29. Voddi, S., Sirisha, U., Praveen, S.P., Sai Pandraju, T.K., Al-Dmour N.A. and Islam, S. Hybrid CNN-GCN Model for Tumor Classification: Integrating Spatial Relationships in Medical Imaging, *2024 International Conference on Decision Aid Sciences and Applications (DASA)*, Manama, Bahrain, **2024**, pp. 1-6.

30. Swapna, D., Sri, U.K., Himaja, V.S.N., Varshita, T.N., Gayatri V. and Praveen, S.P. Crypto Logistic Network: Food Supply Chain and Micro Investment using Blockchain, *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, Pudukkottai, India, **2023**, pp. 908-915.

31. Praveen, S.P., Saripudi, V., Harshalokh, V., Sohitha, T., Venkat Sai Karthik S. and Venkata Pavana Surya Sreekar, T. Diabetes Prediction with Ensemble Learning Techniques in Machine Learning," *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, Pudukkottai, India, **2023**, pp. 1082-1089,

32. Praveen, S.P., Sidharth, S.R., Priya, T.K, Kavuri, Y.S., Sindhura S.M. and Donepudi, S. ResNet and ResNeXt-Powered Kidney Tumor Detection: A Robust Approach on a Subset of the KAUH

Dataset, *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, Pudukkottai, India, **2023**, pp. 749-757.

33. Praveen, S.P., Chaitanya, P., Mohan, A., Shariff, V., Ramesh J.V.N., and Sunkavalli, J. Big Mart Sales using Hybrid Learning Framework with Data Analysis, *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, Pudukkottai, India, **2023**, pp. 471-477.
34. Praveen, S.P., Sindhura, S., Srinivasu P.N. and Ahmed, S. Combining CNNs and Bi-LSTMs for Enhanced Network Intrusion Detection: A Deep Learning Approach," *2023 3rd International Conference on Computing and Information Technology (ICCIT)*, Tabuk, Saudi Arabia, **2023**, pp. 261-268.
35. Srinath Reddy, A., Praveen, S.P., Bhargav Ramudu, G., Bhanu Anish, A., Mahadev A. and Swapna, D. A Network Monitoring Model based on Convolutional Neural Networks for Unbalanced Network Activity, *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, **2023**, pp. 1267-1274.
36. Praveen, S.P., Sarala, P., Kumar, T. N. S. K. M., Manuri, S.G, Srinivas V.S. and Swapna, D. An Adaptive Load Balancing Technique for Multi SDN Controllers," *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, Trichy, India, **2022**, pp. 1403-1409.