ISSN 2007-9737

Comparative Analysis of Classification Models Using Midjourney-generated Images in the Realm of Machine Learning

Anna Karen Gárate-Escamilla[1,*], Rafael Martínez[1], José Carlos Ortiz-Bayliss[1], Amir Hajjam[2]

[1] Instituto Tecnológico de Monterrey,
Escuela de Ingeniería y Ciencias,
Mexico

[2] Université de Technologie de Belfort-Montbéliard,
France

{karen.garate, rafael.mpg, jcobayliss}@tec.mx,
amir.hajjam-el-hassani@utbm.fr

Abstract. Artificial intelligence (AI) integration has shaped rapid and remarkable advances in machine learning. In the relentless pursuit of advancing AI capabilities, applications such as Midjourney emerge as pioneering tools designed to create intricate images from the essence of textual prompts. Midjourney, an example of a generative AI tool, utilizes text-to-image methods with an extensive database. This study aims to provide insights into the potential advantages and limitations of generative AI images in machine learning. This research methodology explores Midjourney to generate 500 images of dogs and cats. Subsequently, these images serve as the basis for building classification models. We will explore the classification models and their evaluations in three scenarios: i) 100% of images generated by Midjourney, ii) 100% of real images, and iii) 50% of images generated by Midjourney and 50% of real images. To achieve this goal, the study utilizes two commonly used deep learning models, InceptionV3 and EfficientNetB4, for training and testing the classification models. The analysis results indicate a significant improvement when combining generated Midjourney images and real images for classification. This comparative examination highlights the effectiveness of AI-generated images in enhancing the performance of machine learning models, emphasizing the potential to augment the image subset with synthesized images from generative IA.

Keywords. Deep learning, neural networks, generative IA, artificial intelligence, machine learning.

1 Introduction

Artificial intelligence (AI) and machine learning (ML) are potent partnerships that have driven significant advances in various fields. At first, this synergy was characterized by discriminative AI, which focused on classification and prediction tasks.

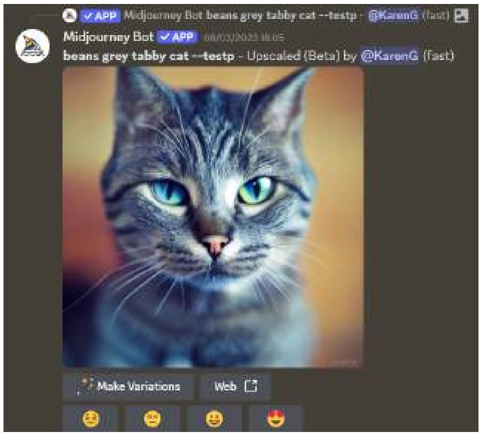
However, an evolving landscape of AI-driven data generation has ushered in a new era with generative AI, capable of generating data from textual inputs [7]. Innovative generative AI applications, notably ChatGPT developed by OpenAI [13], have generated massive attention in advanced AI research.

ChatGPT is known for creating models based on the Generative Pre-trained Transformer (GPT) architecture [19, 14], designed to excel in natural language processing tasks. Another essential part of cutting-edge AI technologies is synthetic image applications, especially Midjourney, DALL-E, and Stable Diffusion, which offer remarkable results by effectively mimicking diverse image content [2].

MidJourney is an innovative AI platform that transforms textual descriptions into visually captivating images using advanced machine learning models.

This involves the interpretation of prompts by a large language model (LLM) and the subsequent

doi: 10.13053/CyS-29-1-5535



✓ APP Midjourney Bot **beans grey tabby cat --testp** - @KarenG (fast)

Midjourney Bot ✓ APP 08/03/2023 18:05

beans grey tabby cat --testp - Upscaled (Beta) by @KarenG (fast)

Fig. 1. Example of a prompt used to generate the picture of a cat in Midjourney

generation of visuals using a diffusion model (DM) trained on a vast corpus of textual and image data. The tool's ability to produce high-quality and artistically appealing images has led to widespread adoption among graphic artists, designers, and other creative professionals [18, 3].

The precision of the text prompts, known as prompt engineering, significantly influences the complexity and quality of the generated images. However, ethical considerations have emerged, including the potential displacement of designers and the unauthorized use of artists' works for training purposes [12].

The main objective is to assess the efficacy and potential benefits of using AI-generated data from Midjourney in enhancing machine learning classification models. Unlike prior research that often considered real and synthetic images in isolation, this study systematically examines the combined effect of real and AI-generated images on the performance of deep learning classification models.

The research specifically focuses on using Midjourney to create a dataset of 500 images featuring dogs and cats. These images are then leveraged to develop and assess classification models in three distinct scenarios: (i) using 100% of images generated by Midjourney, (ii) using 100% of real images, and (iii) combining 50% of images generated by Midjourney with 50% of real ones. For this goal, the study employs two widely recognized deep learning models, InceptionV3 and EfficientNetB4, to train and evaluate the classification models. Our analysis demonstrates an improvement in classification accuracy when combining generated Midjourney images with real images.

This comparison highlights the effectiveness of AI-generated images in enhancing the performance of machine learning classification models, underscoring the potential to enrich the dataset images with synthesized images from generative AI. However, like many AI systems, Midjourney has flaws, and its generated images may exhibit disparities when compared to real-world images.

Consequently, an additional process was required to select the highest-quality images generated by Midjourney. This study provides valuable new insights and understandings within the field of machine learning, examining the potential and limitations of data generated through generative AI.

The remainder of this article is as follows. Section 2 provides the context of our work about related studies. In Section 3, we describe the generation of the synthetic data and the experiment design. Section 4 analyses the results of the experiments conducted. Finally, Section 5 presents the conclusion and suggests potential avenues for future work.

2 Background and Related Work

Artificial intelligence for image generation is a relatively recent topic. However, AI has been used in the past for synthesized image generation. Medel-Vera et al. [8] explored enhancing urban space photographs using generative tools employing convolutional neural networks (CNNs). This approach aims to improve the accuracy and robustness of the classification model by increasing the diversity and quantity of training data using Midjourney and Deep Dream Generator. Ediboglu Bartos et al. [6] conducted a comparative analysis of real and AI-generated synthetic image classification.

They highlight that the selection of the deep learning models, the features, and the task

Table 1. Dog and cat prompts created with the `-testp` command of Midjourney

Image type	Example of prompts
Dogs	"dogs", "full body dog", "dog running, long shot",
	"dog in a park", "dog full body in the sun, blur picture",
	"guardian dog", "dog sleeping",
	"a border collie dog in a beautiful sunset landscape full of grasses and flowers".
Cats	"cats", "full body cat", "cat in the grass",
	"british Shorthair cat", "cat sleeping", "baby cat",
	"cat in a cage", "persian cat, long shot"

influence the model's performance. Baraheem et al. [1] proposed a framework to detect AI-generated images from real ones through CNNs. Their approach achieved perfect accuracy on their test datasets.

Existing research has established a crucial groundwork, yet our work addresses specific aspects and fills gaps identified in previous studies. We aim to use Midjourney to generate synthetic images and integrate them with real images to evaluate their combined impact on machine learning models.

While previous studies tended to compare the performance of models trained only on real or synthetic images, our research goes beyond other works by exploring the hybrid approach of combining 50% real images with 50% AI-generated images. This comprehensive comparison of three different scenarios provides insight into the influence of synthetic data on model training.

This study uses two advanced CNNs: InceptionV3 and EfficientNetB4. Each CNN possesses a unique architectural advantage designed for image classification. InceptionV3, a deep learning model developed by Google, is specifically designed for image classification applications [5, 20].

The critical parts of InceptionsV3 complex architecture include multiple types of convolutional blocks, allowing the network to reduce computation costs through a dimensionality reduction [15].

The inception module involves multiple convolution operations with different kernel sizes and pooling operations, and the outputs are concatenated to the input of the next layer. Our second model, EfficientNetB4, employs a compound scaling method, which scales depth, width, and resolution dimensions.

This model achieves remarkable performance by optimizing the network's architecture using a combination of depthwise convolutions and squeeze-and-excitation layers. This architecture optimization focuses on improving accuracy and computational efficiency, making it a powerful choice for various applications [16].

3 Experimental Design

Our methodology incorporates advanced machine learning algorithms, explicitly emphasizing deep learning models, to assess the efficacy of images generated by generative AI. The ensuing segments delineate the sequential stages of our approach, encompassing the generation of synthetic data, the design of experiments, and the meticulous selection of models.

3.1 Datasets

This research uses Midjourney, a generative AI tool, to generate synthetic images of dogs and cats, which enriches the dataset used for training and testing image classification models. Figure 1 illustrates the result of generating the "beans grey tabby cat" prompt in Midjourney.

Midjourney uses advanced text-to-image techniques to create 500 unique images of dogs and cats based on textual prompts [18].

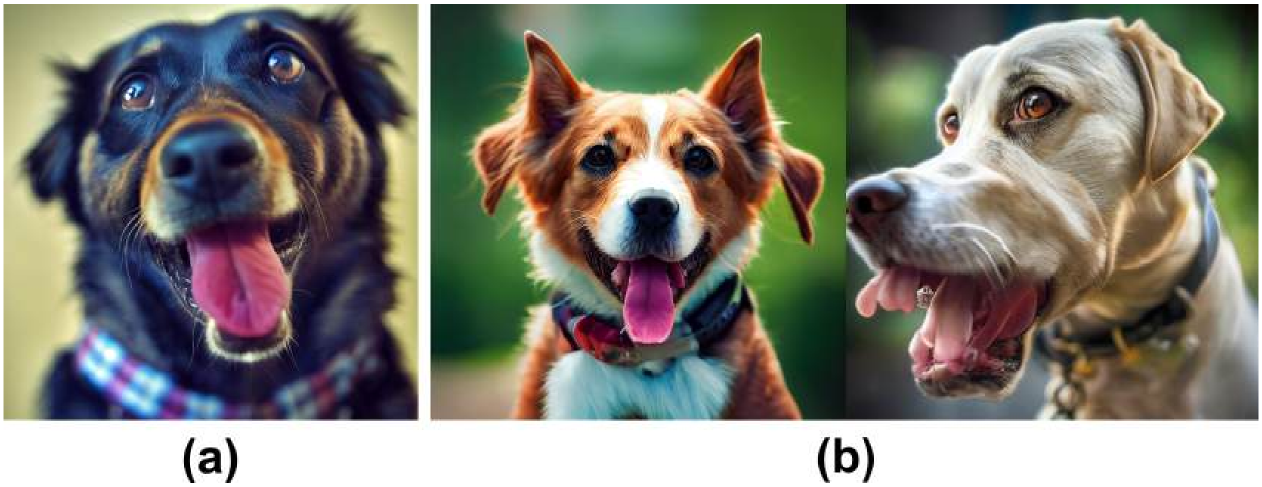


Fig. 2. A comparison of the image quality generated with the prompt "dog with its tongue out". Image (a) shows a more realistic representation of a dog's physical form, while image (b) illustrates distortions such as additional ears or tongues

In addition, we obtained 500 real images of dogs and cats from publicly available datasets [4].

This two-pronged approach ensures a comprehensive dataset that contains both AI-generated and real images, enabling a thorough evaluation of model performance across different types of images. We produced over 2,900 images of cats and dogs with different prompts in Midjourney (from February to May 2023). The prompts were applied with the `-testp` command to give photographic realism to the image [10].

Currently, `-testp` is not available in the v4 and v5 versions of Midjourney, which makes these tests valuable for future comparative analysis. The process of creating synthetic images with Midjourney is quite efficient. Typically, generating a single image takes around 30 seconds to 2 minutes.

Table 1 contains a few examples of prompts used in Midjourney to generate such images. The purpose of the prompts was to generate several real-world scenarios for both cases, including changes in physical locations, different perspectives of the animals, and mentions of some popular breeds of each species.

Since the images created by artificial intelligence can vary in quality, we applied a filtering process to the initial 2900 images we generated with Midjourney to remove images that exhibited evident distortions or unrealistic features.

This included images with anatomical inaccuracies, such as additional ears or misplaced facial features, which did not correspond to the realistic representation of dogs and cats.

After the filtering process, we kept only 500 images. Fig. 2 shows the diversity of possible outcomes when employing the same prompt. As an example, we selected the prompt "dog with its tongue out" to compare the quality and realism of the images produced:

- **Image (a)** provides a more accurate and realistic representation of a dog with its tong out. It focuses on maintaining the correct anatomical features of a real dog. In addition, the image attempts to mimic a photo-realistic camera approach.

- **Image (b)** contains inaccuracies in the correct anatomically correct features of a dog, such as extra ears or tongues. These distortions indicate errors or faults in the image generation process, resulting in an unrealistic portrayal of a dog.

As an additional hurdle for traditional models, the real image dataset is not balanced: As per the repository, the training set has a 75%-25%

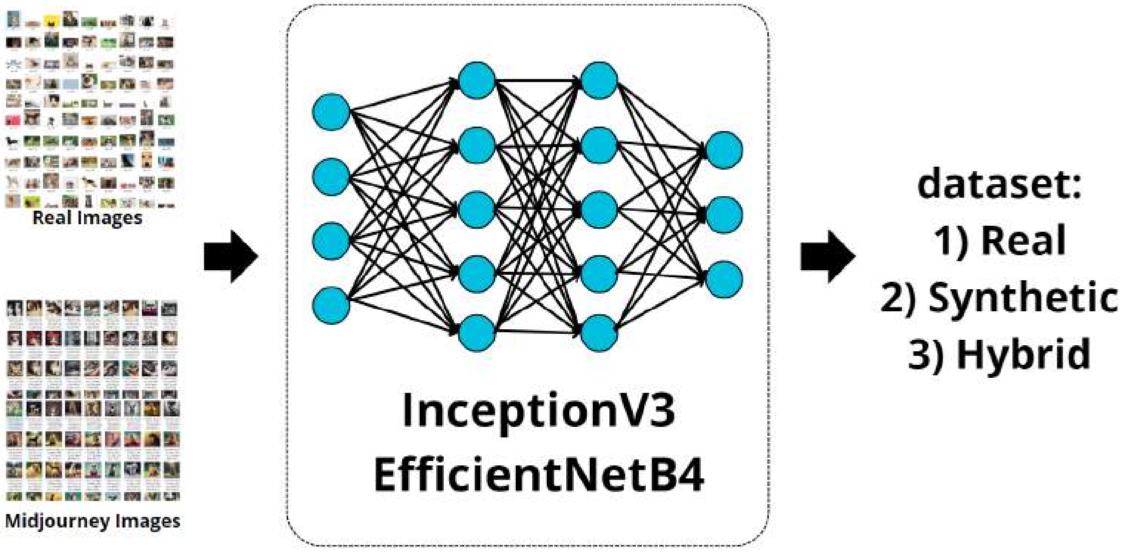


Fig. 3. The pipeline of our experimental design with the scenarios involving the three different datasets

balance of classes, favoring cats. Data imbalance can cause models to place excess importance on a specific class, which harms performance [17].

No further steps are taken to address this imbalance. Instead, we use this as an opportunity to present a more favorable data balance for synthetic data. The Midjourney data has a perfect 50/50 class balance. It should be noted that regardless of the training set used, all models are tested using the same real image test split, which has a 50/50 label occurrence.

Another point to consider is that some data splits have different image amounts. The control dataset contains all 500 real images, the synthetic dataset contains all 500 Midjourney-generated images, and the hybrid dataset contains 1000 images (the whole mix of both).

This is to simulate a real-world scenario where one uses synthetic data to enrich an otherwise limited dataset and to show the feasibility of the data augmentation technique. This also has the secondary effect of slightly reducing class imbalance, as the mixture of frames causes the split to become 60/40 instead of 75/25.

3.2 Experimental Design

The study aims to assess the efficacy of Midjourney-generated images in image classification tasks. It examines three scenarios for evaluation. The first scenario involves training models solely with AI-generated images. The second scenario serves as a baseline, using only real images for comparison.

The third scenario combines both image types, with an equal distribution of 50% Midjourney-generated and 50% real images. The study utilizes two advanced deep learning models, InceptionV3 and EfficientNetB4, for training and testing the classification models.

These models are selected for their established proficiency in image recognition tasks, providing a robust framework for evaluating performance. An overview of our experimental design is illustraaxted in Fig. 3.

3.3 Model Training and Evaluation

The data sets are split into training (80%) and testing (20%) subsets to ensure a reliable assessment of the models. Each model is trained on its respective data set under the three outlined scenarios.

Table 2. A comparison of results between models and data splits. The best results shown in bold. The last column shows the percent change in accuracy compared to the real data split for that model, showing if that split performs better or worse compared to the control. For both models, the best results are for the hybrid splits

Model	Split	Accuracy	IoU	ROC AUC % Change vs Real
InceptionV3	Real	60.00%	70.00%	—
	Synthetic	55.71%	56.90%	-18.71%
	Hybrid	**75.00%**	**84.07%**	**+20.10%**
EfficientNetB4	Real	49.29%	49.29%	—
	Synthetic	**50.00%**	50.00%	+1.44%
	Hybrid	**50.00%**	**52.80%**	**+7.12%**

We acknowledge the importance of using a consistent test partition to evaluate performance in different training scenarios. Specifically, the test set was composed only of real images, which allowed us to evaluate the model performance in classifying real-world data regardless of the composition of the training data.

InceptionV3 and EfficientNetB4 use TensorFlow's Keras API architecture to create the models. These models are configured to include a fully connected classifier layer, accept input images of size 299×299 with three color channels, and do not use pre-trained weights.

We compiled the models with the Adam optimizer and utilized binary cross-entropy as the loss function and a sigmoid activation function for the classifier layer. We evaluated the performance using key metrics such as accuracy and ROC-AUC, providing a comprehensive understanding of the model's effectiveness.

The evaluation seeks to determine the influence of incorporating AI-generated images on classification performance, highlighting any improvements or limitations. Initial findings indicate a significant improvement in model accuracy and generalization when a combination of Midjourney-generated and real images is used, emphasizing the potential of generative AI in enhancing traditional data sets.

4 Results

We summarize our results in Table 2, where the three data splits for both models can be observed.

First, we highlight the control results: InceptionV3 achieved an accuracy of 65% and a ROC-AUC of 70%. As discussed, the test set is balanced, so accuracy can be considered when deciding on the best model.

In general, InceptionV3 performs satisfactorily, though we observed a rapid convergence in the loss during training. We posit that this implies that the model does not have enough data to learn more from the problem — meaning it could benefit from data augmentation techniques.

Before proceeding with the mixed split, we tested the model only on synthetic data. Unsurprisingly, the results were worse: Accuracy of 55.71% and ROC-AU of 56.90%, which the domain shift between synthetic and real data can explain. The distribution is not the same.

Therefore, the model cannot be expected to perform well, yet it undeniably achieved some degree of understanding from utterly fake information. With these two results in mind, two ideas emerge:

– The model would benefit from more data.

– Synthetic data contains information that is of some use to the model.

Finally, we train the model using the hybrid data splits. The performance of the model here (75% of accuracy and 84.07% of ROC AUC) is significantly better, representing a 15% increase in accuracy and a 20% increase in ROC AUC. This shows the feasibility of leveraging the large data pools in generative AI models for data augmentation in the training dataset.

These findings align with previous studies suggesting synthetic data can enhance model performance [9]. To show that these results are not model-specific, we follow the same methodology to train EfficientNetB4. This model, being much less complex than InceptionV3, shows the opposite issue: it fails to learn from the data given.

Though the specific results are worse, we observe that the best split is also the hybrid one, showing a relative improvement of 7.12%. From these results, it is clear that adding AI-generated synthetic data can improve the model's capacity to differentiate between classes when a model lacks data and complexity.

Our research employed advanced synthetic machine learning data augmentation methods to improve the training datasets. In the literature, these deep learning techniques include realistic 3D graphics and image transformation approaches [11]. By generating synthetic Midjouney photographs that closely resemble real-world scenarios, we address the problem of data sparsity within our experimental framework.

5 Conclusion

This paper analyzes the potential benefits and limitations of the generative AI images developed by Midjourney. The research suggests that combining real images with generative AI images improves classification performance, highlighting the viability of image subset augmentation.

When comparing the hybrid model with the real dataset, the InceptionV3 model showed a marked improvement in accuracy and ROC-AUC. Although EfficientNetB4, had a smaller benefit, a considerable improvement can be observed against the real data.

In a broader context and applicability, this study highlights the potential of generative AI to complement real images by improving the metrics of machine learning models.

Thus, the results suggest that incorporating generative AI images can augment datasets, mainly when real photographs may be limited in some cases, such as specialized areas and camera limitations. This can be seen in specialized domains such as veterinary medicine and wildlife monitoring, where obtaining high-quality images can be challenging and resource-intensive. Our study provides valuable insights related to generative AI for image classification.

However, we have identified some points that we should consider as future research. The number of available photographs constrains the current study, and additional image generation would improve the dataset's quality.

We focused on InceptionV3 and EfficientNetB4, and while these models produced promising results, exploring additional model architectures would benefit our analysis. Expanding the dataset with diverse and complex photographs will also be essential for future research. Furthermore, it would be advantageous to explore other generative AI tools to compare their effectiveness.

References

1. **Baraheem, S. S., Nguyen, T. V. (2023).** AI vs. AI: Can AI detect AI-generated images?. Journal of Imaging, Vol. 9, No. 10, pp. 199. DOI: 10.3390/jimaging9100199.
2. **Borji, A. (2022).** Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and DALL-E 2. DOI: 10.48550/ARXIV.2210.00586.
3. **Chen, H. C., Chen, Z. (2023).** Using ChatGPT and Midjourney to generate chinese landscape painting of tang poem 'The difficult road to Shu'. International Journal of Social Sciences and Artistic Innovations, Vol. 3, No. 2, pp. 1–10. DOI: 10.35745/ijssai2023v03.02.0001.
4. **Cortinhas, S. (2024).** Cats and dogs image classification. www.kaggle.com/datasets/samuelcortinhas/cats-and-dogs-image-classification.
5. **Dong, N., Zhao, L., Wu, C., Chang, J. (2020).** Inception v3 based cervical cell classification combined with artificially extracted features. Applied Soft Computing, Vol. 93, pp. 106311. DOI: 10.1016/j.asoc.2020.106311.

6. **Ediboglu-Bartos, G., Özmen-Akyol, S. (2023).** Deep learning for image authentication: A comparative study on real and AI-generated image classification. Proceedings of the Annual Insights Symposium 18th International Symposium on Applied Informatics and Related Areas, pp. 1–5.

7. **Lv, Z. (2023).** Generative artificial intelligence in the metaverse era. Cognitive Robotics, Vol. 3, pp. 208–217. DOI: 10.1016/j.cogr.2023.06.001.

8. **Medel-Vera, C., Vidal-Estévez, P., Mädler, T. (2024).** A convolutional neural network approach to classifying urban spaces using generative tools for data augmentation. International Journal of Architectural Computing, Vol. 22, No. 3, pp. 392–411. DOI: 10.1177/14780771231225697.

9. **Meiser, M., Zinnikus, I. (2024).** A survey on the use of synthetic data for enhancing key aspects of trustworthy AI in the energy domain: Challenges and opportunities. Energies, Vol. 17, No. 9, pp. 1992. DOI: 10.3390/en17091992.

10. **Midjourney (2024).** Legacy model version parameters. docs.midjourney.com/docs/legacy-model-parameters.

11. **Mumuni, A., Mumuni, F., Gerrar, N. K. (2024).** A survey of synthetic data augmentation methods in machine vision. Machine Intelligence Research, Vol. 21, No. 5, pp. 831–869. DOI: 10.1007/s11633-022-1411-7.

12. **Olaoye, F., Potter, K. (2024).** Ethical considerations in artificial intelligence. EasyChair.

13. **OpenAI (2024).** Chatgpt. chatgpt.com/.

14. **Ray, P. P. (2023).** ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems, Vol. 3, pp. 121–154. DOI: 10.1016/j.iotcps.2023.04.003.

15. **Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016).** Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826. DOI: 10.1109/cvpr.2016.308.

16. **Tan, M., Le, Q. V. (2020).** EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36 th International Conference on Machine Learning, pp. 1–11.

17. **Thabtah, F., Hammoud, S., Kamalov, F., Gonsalves, A. (2020).** Data imbalance in classification: Experimental evaluation. Information Sciences, Vol. 513, pp. 429–441. DOI: 10.1016/j.ins.2019.11.004.

18. **Yin, H., Zhang, Z., Liu, Y. (2023).** The exploration of integrating the midjourney artificial intelligence generated content tool into design systems to direct designers towards future-oriented innovation. Systems, Vol. 11, No. 12, pp. 566. DOI: 10.3390/systems11120566.

19. **Zhong, Q., Ding, L., Liu, J., Du, B., Tao, D. (2023).** Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT. DOI: 10.48550/ARXIV.2302.10198.

20. **Zhou, W., Wang, H., Wan, Z. (2022).** Ore image classification based on improved CNN. Computers and Electrical Engineering, Vol. 99, pp. 107819. DOI: 10.1016/j.compeleceng.2022.107819.

Article received on 28/05/2024; accepted on 04/07/2024.
**Corresponding author is Anna Karen Gárate-Escamilla.*