JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Exploring AI Image Generation: A Comparative Insights from API Integration

Dr. Snehal Rathi
Vishwakarma Institute of Information Technology, Pune, Maharashtra, India
Email:snehal.rathi@viit.ac.in

Darshan Kotecha
Vishwakarma Institute of Information Technology, Pune, Maharashtra, India
Email: *darshan.22110342@viit.ac.in*

Sahil Savardekar
Vishwakarma Institute of Information Technology, Pune, Maharashtra, India
Email: *sahil.22110350@viit.ac.in*

Gunjan Ghuguskar
Vishwakarma Institute of Information Technology, Pune, Maharashtra, India
Email: *gunjan.22220021@viit.ac.in*

Vaibhav Sawate
Vishwakarma Institute of Information Technology, Pune, Maharashtra, India
Email: *vaibhav.22110326@viit.ac.in*

***Abstract* - The most difficult jobs in computer vision is creating a picture from various data formats, such as text, scene graphs, and object layouts. Furthermore, manually taking pictures from various angles in order to create an object or a product can be time-consuming and complex. Deep learning and artificial intelligence algorithms have now made it feasible to create fresh images from various kinds of data. For this reason, a lot of work has recently been put into creating image-generating techniques, with remarkable success. In light of this, we provide a thorough analysis of current picture creation techniques in our study, to the best of our author's knowledge. As a result, an explanation of each image creation method is carried out according to the primary goal, the type of data utilized, and the nature of the employed algorithms. Additionally, the suggested methods are shown in order to discuss each image-generating category. A live implementation of the DALL-E model by ChatGPT is also been implemented. To further explain the latest advancements and pinpoint strengths and limits, a comparing performance of current solutions is presented together with a discussion of the assessment metrics appropriate for each picture-generating category. Finally, the issues that this topic is currently facing are discussed. Here, in this paper we deal with creating a picture from simple "text" data formats and prompts. The selected method makes it simpler to create images from user-provided text inputs by combining many pre-trained models with expertise in computer vision and natural language processing (NLP). Simplifying the process of creating unique photos and creating new opportunities for advertising, content, and art are the objectives. While guaranteeing the accuracy and quality of the information produced, the study also investigates the performance and adaptability of several AI models.**

Keywords: Gemini, Bing, Copilot, OpenAI, Transformers, Diffusion, Neural Networks, GAN, VAE, BERT, AutoEncoders, React

I. INTRODUCTION

Artificial intelligence (AI) image generation technologies convert word descriptions or other input data into realistic, high-quality images using intricate algorithms. This feature has transformed industries that require automated and changeable images, such marketing, entertainment, and content production. These AI systems rely on a number of methods, including Diffusion models, Transformer-based architectures, and Generative Adversarial Networks (GANs). This study looks at these algorithms and how they are used in well-known AI platforms. It then compares them to show their advantages, disadvantages, and potential uses. Properly labeled data that allows information to be taken from the images for numerous inferences is crucial to the AI learning process. Therefore, discovering automatically linked material and automatically annotating it has become a serious problem for AI techniques in various tasks. The growth of AI and an ability of computers to study and generate fresh images by processing large datasets from several taught scenarios, which is a good option. Creating fresh images is useful for a variety of additional applications, including fugitive detection, data augmentation, and object reconstruction. One of the most difficult jobs in computer vision is creating a new image from an existing one. The issue of creating fresh images with excellent performance in terms of image quality and relevant information was resurrected a few years ago, with the advent of AI techniques such as generative adversarial networks (GANs). It has created lifelike pictures of people, objects, or landscapes that are hard to tell apart from authentic photos.

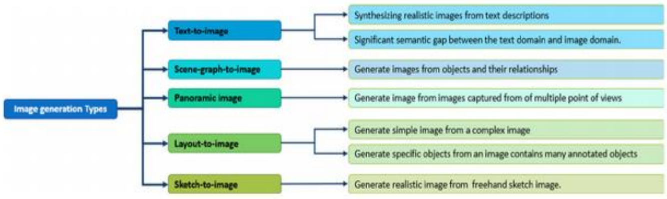


Image generation Types
Text-to-image
Synthesizing realistic images from text descriptions
Significant semantic gap between the text domain and image domain.
Scene-graph-to-image
Generate images from objects and their relationships
Panoramic image
Generate image from images captured from of multiple point of views
Layout-to-image
Generate simple image from a complex image
Generate specific objects from an image contains many annotated objects
Sketch-to-image
Generate realistic image from freehand sketch image.

Fig.1 Image Generation ways based on user input

Consequently, various techniques are utilized to produce specific images through AI models. These techniques include picture-to-image translation, sketch-to-image generation, conditional image generation, text-to-image generation, video generation, panoramic image generation, and scene-graph image generation.

II. LITERATURE SURVEY

A subset of generative models, picture generative AI models seek to produce realistic and cohesive images from scratch. These models extract patterns and characteristics from a large amount of training data using sophisticated algorithms and deep learning approaches. They fall into one of two general categories:

Variational Autoencoders (VAEs): These probabilistic models encode pictures into vector representations in a latent space. The model can then create new images by drawing samples from the latent space after the decoder reconstructs the images from the encoded vectors. (Shown in Fig.2)

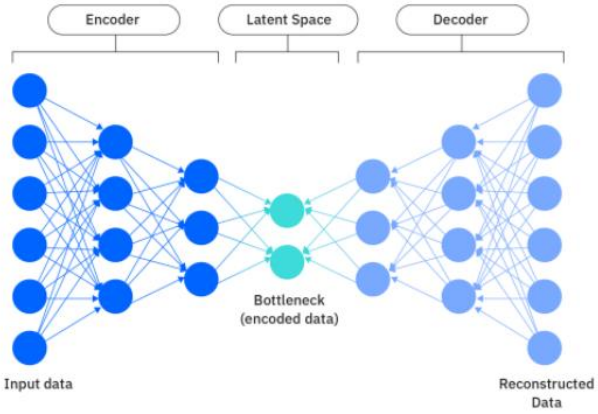


Fig.2 A visual depiction of autoencoder neural network

Generative Adversarial Networks (GANs): It consist of a generator, and a discriminator, as these two neural networks (4609-4646) (shown in Fig.3) engaged in a competitive process. In order to deceive the discriminator, which seeks to differentiate between authentic and fraudulent images, the generator produces artificial images. (Elasri, 2022) The outcome of this back-and-forth conflict is the creation of incredibly lifelike visuals.[1]

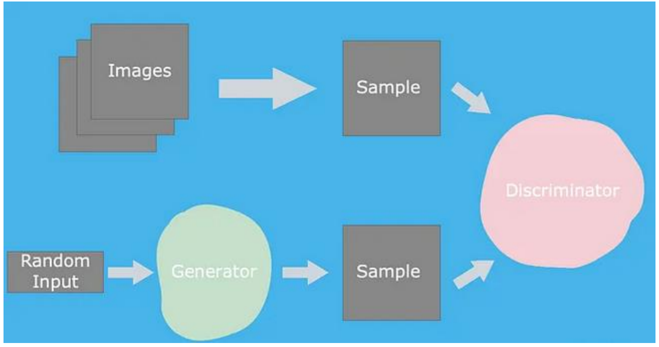


Fig.3 Image Generation types based on user input

Similar to VAEs, GANs are a joint architecture that combines two neural networks: a network that determines whether a particular image is a "fake" image from the generator network or a "real" image from the training data set, and a generator network that produces image samples that resemble images from the training data set. (Yadav, (2023).)[2]

Diffusion Model: Diffusion models are a newer approach in the field of image generation, wherein the model gradually transforms random noise into coherent images. This process involves multiple iterations of refinement, resulting in high-quality, detailed visuals.
It has been demonstrated that Diffusion-models, a relatively new kind of generative model, can produce realistic visuals. Additionally, they frequently train more quickly than other generative model types like GANs. Process is as follows:

1. Generate an image of noise with a high entropy level.
2. Use a diffusion process to add detail to the noise image.
3. Use a neural network to control the diffusion process.
4. Generate a realistic image by running the diffusion process until the required detail is achieved. (shown in Fig.4)

Gemini uses **BERT (Bidirectional Encoder Representations from Transformers)** architecture for natural language processing and generation. BERT doesn't generate text itself but does tasks that require a thorough understanding of language. Think questioning, sentiment analysis, and summarization. For instance, having to rapidly understand the main ideas of a long study paper or figuring out how a customer review feels overall. BERT is capable of analyzing the language and giving you the necessary final results.

The ability of Gemini to digest and comprehend information from several modalities with ease is what sets it apart. Gemini can do a bigger variety of jobs with increased accuracy and efficiency thanks to their multimodal prowess (Ouyang, 2022).[7] **Using Multimodal Processing** Gemini can effortlessly shift among text, code, images, and videos, allowing it to establish links across these diverse data sources. This capability is crucial for activities such as medical diagnosis, where comprehending patient information, imaging, and clinical notes is vital. **Gemini's Seamless Tool Integration** allows for easy integration with existing software and tools. This feature makes it easier to integrate Gemini's powerful capabilities into other applications and workflows.[3]

Microsoft Copilot, integrated within the Microsoft 365 suite, uses large language models (**LLMs**) based on transformer architectures, similar to those powering OpenAI's models. Although primarily a text-focused system, Copilot can handle visual data by leveraging text-to-image modules (e.g., DALL-E APIs) for image generation tasks. This integration enables it to enhance productivity tools by automatically generating images that align with document content.

DALL·E is the first generative pre-trained transformer (GPT) model using transformer architecture. It comprises three components:
1. Discrete **VAE**

2. An autoregressive decoder-only **transformer**
3. A CLIP pair of **image-encoder** and **text-decoder**.

It enables user to create image via text prompts, It is trained neural network that fetch and generate images in variety of images.

The transformer employs BPE encoding that accepts text and images as a unified data stream holding a maximum of 1280 tokens (256 for text and 1024 for images), modelling them all autoregressively and trained with maximum chances to produce each token in sequential order. As per the text guidelines, this training method allows DALL·E to both generate an image from its raw context and to reconstruct any a rectangular section of the current image that reaches the bottom-right corner. (Zhu, 2024)[10]

Each image token can focus on each text token due to the attention mask present in all 64 of its self-attention layers. Based on the layer, DALL·E utilizes a row, column, or convolutional attention scheme for the image tokens and a causal mask for the text tokens.

It can generate "variations" of the image as separate outputs derived from the original, and also alter the image to change or enhance it.

The datasets for training DALL-E consist of images from Wikipedia and YFCC100M++. Minor captions, foreign languages, dates, and unusual aspect ratios are some of the criteria applied to clean the data. The datasets CUB-200 and MS-COCO are utilized for testing purposes.

III. METHODOLOGY:

An intriguing utilization of image generative AI models is text-to-image synthesis, allowing the transformation of written descriptions into matching visual depictions. This approach entails integrating natural language processing (NLP) (Ouyang, 2022)[7]methods in with image generation models to obtain remarkable outcomes.

Conditional GANs for Text-to-Image: The integration of text embedding and GANs has been successfully achieved, resulting in the generation of images that are based on specific textual input. This suggests that the model can generate exceptionally accurate and intricate images when provided with a comprehensive textual description.

Image Synthesis: Text-to-image technology is utilized across multiple sectors, such as virtual reality, e-commerce, and content creation. For instance, this technology can accelerate product design and development in e-commerce by generating product visuals from written descriptions.

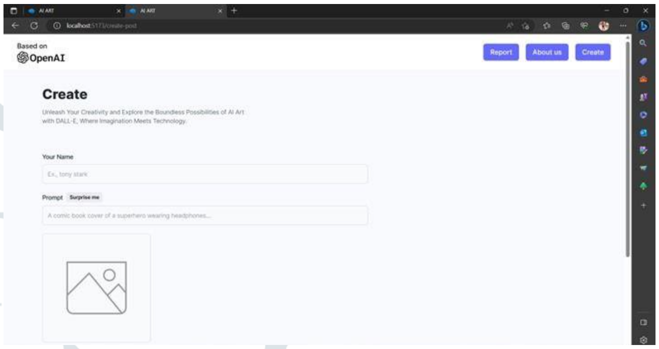
In this project we have integrated the DALL-E API in our react based web application. To ensure a smooth user experience, the integration of this model involved several key steps:

Secure API access was established using authentication keys provided by the platform (i.e, OpenAI DALL-E). This ensures safe communication between the platform and the models. After successful integration, when a user submits a text description, it is formatted and sent as a request to the selected AI model's API. Users can provide varying levels of detail in their prompts, allowing for flexibility in image output. After the text input is processed, the API returns an image in a standardized format (JPEG or PNG). The platform then displays the image, and basic post-processing is applied as necessary. So the model optimize the generated images for different purposes using various available methods like resizing and image enhancements like clarity, contrast etc.

MongoDB is our cloud database in this project. Cloud databases are similar to traditional databases, but they do not require any setup or infrastructure maintenance. Cloud databases are hosted in a cloud computing environment.

IV. RESULTS & DISCUSSION

In this project we have used Visual Studio Code for programming.
1. We have used Cloud MongoDB as our database.
2. The Front-end is implemented in NodeJS.
3. The Back-end is implemented using ReactJS



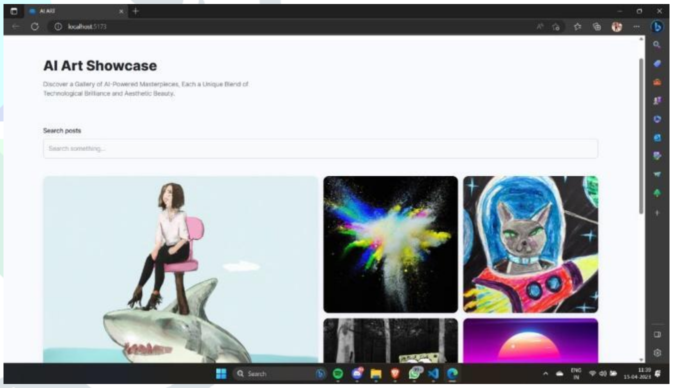
Create

Unleash Your Creativity and Explore the Boundless Possibilities of AI Art with DALL-E, Where Imagination Meets Technology.

Your Name: Ex., tony stark

Prompt: A comic book cover of a superhero wearing headphones...

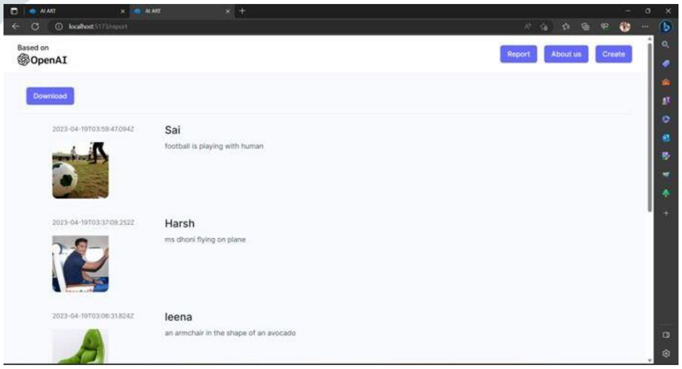
Fig.8 User friendly frontend



AI Art Showcase

Discover a Gallery of AI-Powered Masterpieces, Each a Unique Blend of Technological Brilliance and Aesthetic Beauty.

Search posts: Search something...

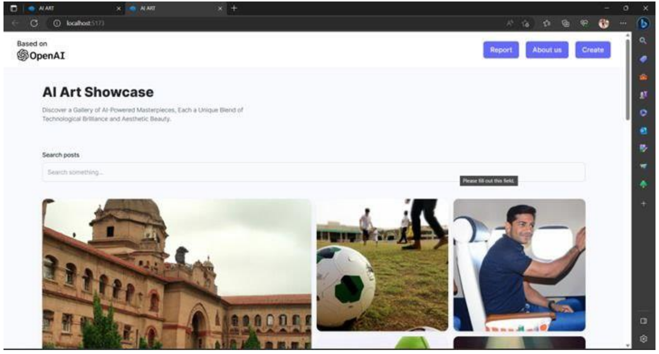


2023-04-19T03:59:47.094Z Sai
football is playing with human

2023-04-19T03:37:09.252Z Harsh
ms dhoni flying on plane

2023-04-19T03:06:31.824Z leena
an armchair in the shape of an avocado

Fig.9 Web-application with OpenAI API Integration



AI Art Showcase

Discover a Gallery of AI-Powered Masterpieces, Each a Unique Blend of Technological Brilliance and Aesthetic Beauty.

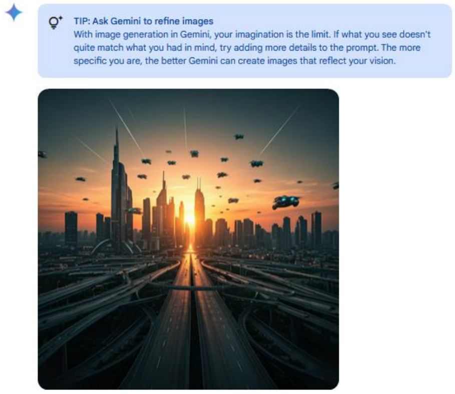
Search posts

Search something...

Please fill out this field.

Fig.10 Stored database

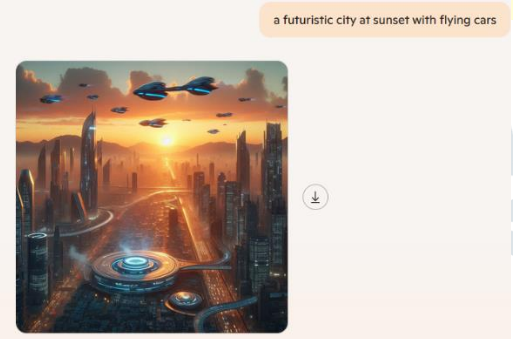
Here, lets start with the prompt "a futuristic city at sunset with flying cars" and check the comparison in particular outputs.



TIP: Ask Gemini to refine images

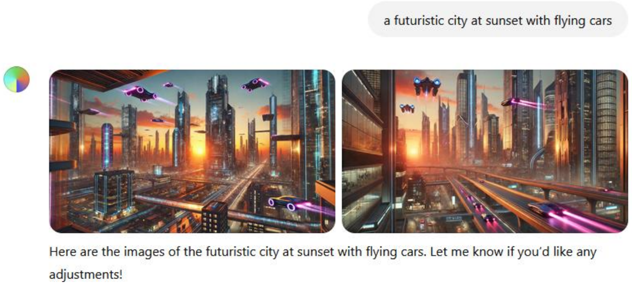
With image generation in Gemini, your imagination is the limit. If what you see doesn't quite match what you had in mind, try adding more details to the prompt. The more specific you are, the better Gemini can create images that reflect your vision.

Fig.11 Image Generated by Gemini



a futuristic city at sunset with flying cars

Fig.12 Image Generated by Copilot



a futuristic city at sunset with flying cars

Here are the images of the futuristic city at sunset with flying cars. Let me know if you'd like any adjustments!

Fig.13 Image Generated by DALL.E based on GPT-3

Here, it can be observed that the ability to generate images based on various prompts and styles.

Although Gemini and OpenAI's GPT-3 have certain similarities, Gemini's real strength is its capacity to process and produce meaningful text in addition to visual output. Although it is capable of producing text, GPT-3 is not naturally able to understand and create material that combines textual and visual modes.

V. CONCLUSIONS:

DALL·E 2's comprehension of language is restricted. It can be challenging to differentiate between "A black car, a white bike, A red bike, and a black car, as well as between "An artist creating sand art" and "Sand art created by an artist." It creates visuals of "a cowboy on a horse" when given the prompt "a cowboy riding a horse." It also does not produce the accurate images in different situations. Asking for over three items, using negation, numbers, and linked sentences can lead to errors, causing features of objects to be attributed incorrectly. Additional limitations consist of its restricted capability to manage scientific information, like astronomy or medical images, and its handling of language, which almost invariably leads to surreal nonsense despite using recognizable characters.

This comparative analysis shows the diversity in AI image generation technologies and highlights how different platforms leverage distinct algorithms to meet specific user needs.

Parameter	Gemini (Google)	Microsoft Copilot	ChatGPT (OpenAI)
Primary Algorithm	Transformer-based, multimodal	Transformer with text-to-image integration	Transformer with multimodal capabilities
Image Quality	High-quality, flexible	Variable (depends on API integration)	High-quality, contextually relevant
Versatility	Text, image, and multimodal tasks	Text and image (via integration)	Primarily text, with image generation
Computational Efficiency	Moderate to high	High (text-focused with image as add-on	High; uses cloud for multimodal tasks
Ease of Integration	API access, adaptable to multiple platforms	Integrated in Microsoft 365 tools	Integrated in ChatGPT for multimodal use
Application Areas	Creative content, multimedia	Document productivity, visual aids	Conversational, educational, creative
Limitations	Limited by API access policies	Limited creative flexibility	Limited detailed image control

VI. REFERNCES:

1. Elasri, Mohamed, Omar Elharrouss, Somaya Ali Al-Maadeed and Hamid Tairi. "Image Generation: A Review." *Neural Processing Letters* 54 (2022): 4609-4646.
2. Yadav, Archana. (2023). GEN AI-DRIVEN ELECTRONICS: INNOVATIONS, CHALLENGES AND FUTURE PROSPECTS. 10.5281/zenodo.8165255.
3. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb), 1137–1155.
4. Zhang, X., Liu, S., Zhang, R., Liu, C., Huang, D., Zhou, S., Guo, J., Kang, Y., Guo, Q., Du, Z., & Chen, Y. (2020). Adaptive precision training: Quantify backpropagation in neural networks with fixed-point numbers. *arXiv preprint arXiv:1911.01989*. https://arxiv.org/abs/1911.01989

5. Poole, B., Jain, A., Barron, J. T., & Mildenhall, B. (2022). DreamFusion: Text-to-3D using 2D diffusion. *arXiv preprint arXiv:2209.14916*. https://arxiv.org/abs/2209.14916
6. Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., & Kulshreshtha, A. (2022). LaMDA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*. https://arxiv.org/abs/2201.08239
7. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Al-Rfou, R., Choe, D., Constant, N., Guo, M., & Jones, L. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*. https://arxiv.org/abs/2203.02155
8. Al-Rfou, R., Choe, D., Constant, N., Guo, M., & Jones, L. (2018). Character-level language modeling with deeper self-attention. *arXiv preprint arXiv:1808.04444*. https://arxiv.org/abs/1808.04444
9. Ding, N., Qin, Y., Yang, G., et al. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence, 5,* 220–235. https://doi.org/10.1038/s42256-023-00626-4
10. Zhu, L., Lai, Y., Mou, W., Zhang, H., Lin, A., Qi, C., Yang, T., Xu, L., Zhang, J., & Luo, P. (2024). ChatGPT's ability to generate realistic experimental images poses a new challenge to academic integrity. *Journal of Hematology & Oncology, 17*. https://doi.org/10.1186/s13045-024-01543-8
11. "Exploring the Impact of Different Interaction Design Principles on User Experience: Case Studies and Experiments", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.11, Issue 5, page
12. S. Rathi, P. Wawage and A. Kulkarni, "Automatic Question Generation from Textual data using NLP techniques," 2023 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2023, pp.1-7, doi: 10.1109/ESCI56872.2023.10100278.
13. 25. S. P, S. R. Rathi, M. Singh, P. Naval, K. Batri and B. Agnihotri, "An Exploration of the Use Cases for Artificial Intelligence in Data Access Optimization," 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Bangalore, India, 2023, pp. 1-6, doi:10.1109/SMARTGENCON60755.2023.10442842.
14. S. Rathi, P. Chate, G. Desai, O. Gangji, V. Kale and A. Kalbhor, "Cross-Modal Question Generation: NLP-based Approaches for Text, Image, PDF, and Video Inputs," 2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bengaluru, India, 2023, pp. 992-1002, doi: 10.1109/ICIMIA60377.2023.10426142.