Narrative Ai: Visual Storycraft

Mr. Aneesh Kumar A[1], Hemanth B[2], Sharath M D[3], Prajwal D[4], Mohammad Wafeeq[5]
[1] *Assistant Professor, Dept. of CSE (IoT & Cyber Security with Blockchain Technology), Mangalore Institute of Technology & Engineering, Moodabidri, India*
[2,3,4,5] *Student, Dept. of CSE (IoT & Cyber Security with Blockchain Technology), Mangalore aInstitute of Technology & Engineering, Moodabidri, India*

Abstract: This project leverages artificial intelligence to create interactive and dynamic storytelling from images. By combining image captioning, language models, and text-to-speech technologies, the system generates engaging narratives based on the content of uploaded images. Using the BLIP image captioning model, the project converts visual content into descriptive text. This description is then processed by the Falcon-7B-Instruct language model to generate a short story. The final output is a text-to-speech conversion of the story, offering a fully automated, multimedia storytelling experience. The system is built using Python, Streamlit, and Hugging Face APIs, providing users with an easy to-use web interface to upload images, generate stories, and listen to the narratives.

Keywords: Artificial intelligence, interactive storytelling, dynamic storytelling, image captioning, language models, text-to-speech (TTS), BLIP model, Falcon-7B-Instruct model, descriptive text generation, short story creation, multimedia experience, Python, Streamlit, Hugging Face APIs, web interface, and narrative generation.

I. INTRODUCTION

Narrative AI: Visual Storycraft is an advanced AI-powered project designed to bridge the gap between visual and linguistic intelligence, enabling the transformation of images into captivating narratives accompanied by audio narration. This tool leverages cutting-edge AI technologies, combining image captioning, natural language processing, and text-to-speech synthesis to create an immersive storytelling experience. The process begins with analyzing the visual content of an uploaded image using the HuggingFace Salesforce/blip-image-captioning-base model, which generates a detailed textual description. This description then serves as the foundation for story generation, utilizing the tiiuae/falcon-7b-instruct large language model to craft creative and coherent narratives. Finally, the generated story is converted into high-quality audio using the HuggingFace espnet/kan-bayashi_ljspeech_vits text-to-speech model, offering users the ability to listen to the stories directly through the application.

The system is built with Python and Streamlit, providing an intuitive and interactive web interface where users can upload images, view the descriptive text and narrative, and play the audio. This tool serves diverse applications across creative writing, education, entertainment, and accessibility. Writers and storytellers can use it for inspiration, educators can engage students with narrative-driven learning, and parents can create personalized stories for children. Moreover, it promotes accessibility by enabling visually impaired individuals to experience the richness of visual content through audio storytelling. By integrating creativity and automation into a unified platform, Narrative AI: Visual Storycraft addresses the growing demand for tools that seamlessly combine image, text, and speech modalities, empowering users of all backgrounds to engage with and create compelling narratives effortlessly.

II. LITERATURE SURVEY

The existing systems for automatic story generation exhibit various limitations in creativity, adaptability, and contextual understanding. Traditional approaches, as discussed by Luís Miguel Botelho, rely on predefined templates and rule-based methodologies, resulting in logical but monotonous and rigid storytelling that lacks diversity and engagement. Similarly, studies by Holy Lovenia et al. highlight the shortcomings of image-to-story systems that focus on translating visual content into textual descriptions without incorporating stylistic diversity, emotional depth, or thematic context, leading to generic outputs. Ali Farhadi et al. emphasize that conventional image captioning systems generate factual descriptions that fail to form coherent narratives, lacking the ability to interpret complex visual scenes or contextualize

relationships. Further, as noted by Sonali Fotedar et al., traditional storytelling systems depend on predefined scripts and struggle to adapt to varying genres, thematic nuances, or audience preferences, often resulting in rigid and predictable narratives. Lastly, Mariët Theune et al. point out that earlier narrative generation models, rooted in logical frameworks or rule-based methods, produce outputs that are generic and predictable, with minimal creativity and limited capacity for simulating complex character interactions or evolving plot structures. Collectively, these studies underscore the need for innovative, flexible, and context-aware systems capable of generating engaging and personalized narratives.

III. SCOPE AND METHODOLOGY

Scope

This work focuses on developing and evaluation of a framework based on deep learning so that detection and classification in skin cancer can be done, keeping in view the need gap of the current diagnostic practices. The proposed framework uses sophisticated convolutional neural networks and applies transfer learning to analyze massive dermatological datasets, bypassing problems such as balancing of data using augmentation techniques. This framework aims to increase the diagnostic accuracy while including diverse demographics that can be interpreted by the attention mechanisms like Grad-CAM. It is scalable in design and supports early detection as well as resource optimization for areas which are underserved. In all these research, ethical concerns are maintained that have priority on mitigating biases, privacy of data, and integrating smoothly with clinical workflows that, at last, lead to fair and effective dermatological care.

Methodology

The methodologies employed in existing systems for automatic story generation demonstrate a progression in integrating advanced computational techniques with creative storytelling. Luís Miguel Botelho's system emphasizes semantic modeling and machine learning to simulate complex character interactions and evolving plot structures, combined with creative heuristics and linguistic tools for thematic richness and coherence. Holy Lovenia et al. propose a multimodal framework that utilizes Convolutional Neural Networks (CNNs) and Transformers to extract image features and generate stylistically controlled narratives, enabling user-defined tonal adjustments. Ali Farhadi et al. focus on aligning visual and textual embeddings through dual-encoder frameworks, leveraging CNNs and Recurrent Neural Networks (RNNs) or Transformers to create descriptive sentences grounded in semantic segmentation. Sonali Fotedar et al. incorporate Transformer-based architectures, such as GPT, alongside reinforcement learning to ensure thematic consistency and optimize story arcs, blending predefined templates with generative algorithms for adaptability across genres. Mariët Theune et al. integrate natural language generation with narrative planners to structure storylines and character arcs, balancing rule-based and statistical methods to enhance creativity and coherence. Across these methodologies, common elements include the use of multimodal AI techniques, iterative refinement through feedback mechanisms, and modular designs to support scalability and adaptability, collectively addressing the challenges of generating diverse, engaging, and contextually relevant narratives.

IV. SYSTEM ARCHITECTURE

The system design of the Narrative AI project is structured into three primary layers—Front-End Layer, Back-End Layer, and AI Models Layer—which work together to deliver a seamless and efficient user experience, with cloud integration ensuring scalability and reliability. The Front-End Layer provides an interactive user interface where users upload images in formats like JPG. After an image is uploaded, the system processes it using the BLIP model to generate a descriptive caption, which is displayed as the "scenario." This caption is then sent to the Falcon-7B model, which creates a detailed story based on the description, and the generated story is displayed on the interface. Additionally, the ESPNet text-to-speech (TTS) model converts the story into audio, which is played back to the user via an integrated audio player, offering both visual and auditory storytelling experiences.

The Back-End Layer manages the core processing and logic of the system. Upon receiving the uploaded image, it interacts with the BLIP model to generate a description. This description is passed to the Falcon-7B model for story generation, which processes the text and generates a coherent narrative.

The story is then sent to the ESPNet TTS model to convert it into audio. API calls are used to facilitate communication with external models hosted on platforms like HuggingFace, ensuring that the system operates efficiently.

The AI Models Layer is composed of three pre-trained models that handle the image processing, story generation, and audio conversion tasks. The BLIP model converts images into descriptive text, which serves as the foundation for the story. The Falcon-7B model generates the story from the image caption, ensuring it is creative and contextually relevant. Finally, the ESPNet model converts the generated story text into audio, enhancing the user's experience with an immersive auditory output.

The system follows a straightforward data flow: first, the user uploads an image, which is then processed by the BLIP model to generate a description. This description is sent to the Falcon-7B model to create the story, and the generated story is passed to the ESPNet model to produce an audio file. Both the text and audio are returned to the front-end, where users can view the story and listen to the narrative. To support the system's scalability and ensure that it can handle multiple user requests and real-time processing, the back-end is hosted on cloud services such as AWS, Google Cloud Platform, or Heroku. The AI models themselves are hosted on platforms like HuggingFace, accessed via API calls, ensuring efficient and seamless integration for processing image-to-text, story generation, and text-to-speech conversion.

This multi-layered architecture enables the system to provide a holistic, user-friendly solution for transforming images into dynamic and engaging stories, offering both visual and audio component.

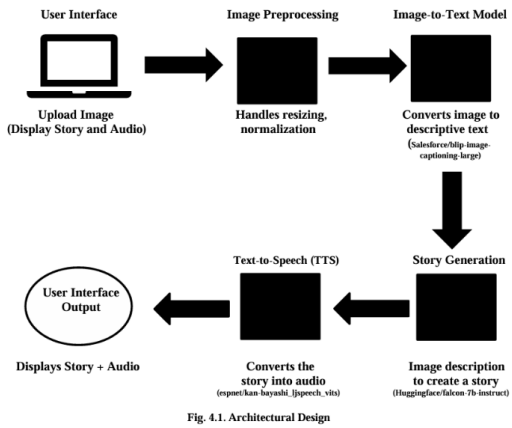

Figure 1: System architecture

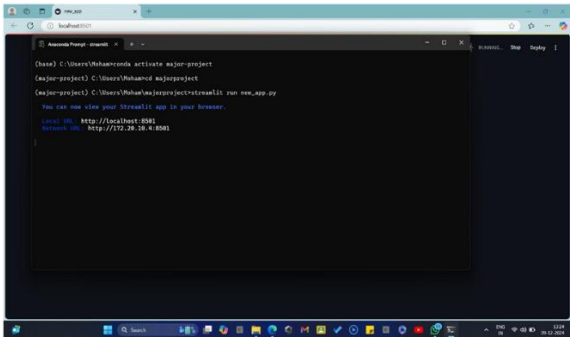

Figure 2: Output1

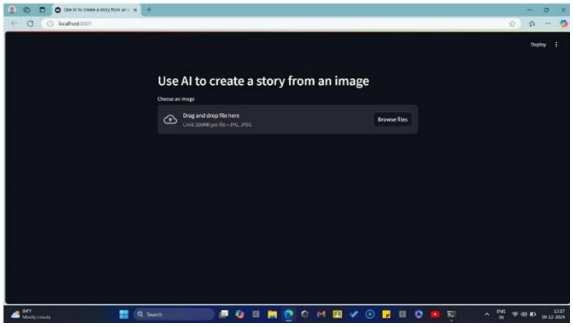

Figure 3: Output 2

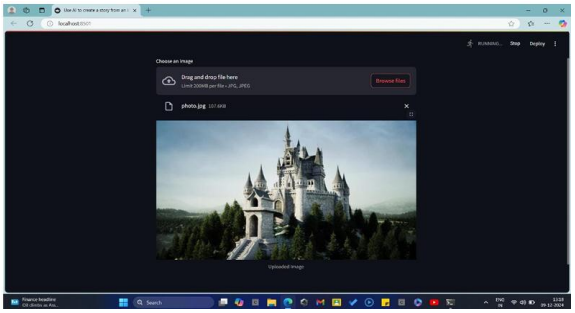

Figure 4: Output 3

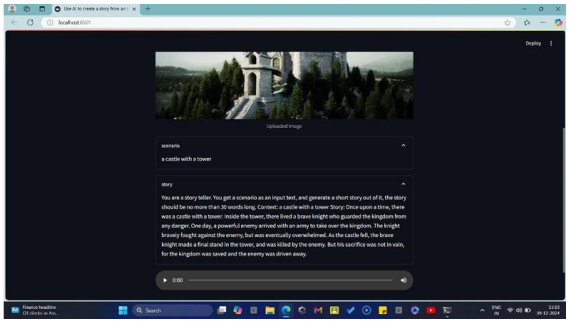

Figure 4: Output 3

V. CONCLUSION

In conclusion, the Narrative AI project offers an innovative solution for transforming images into dynamic stories with accompanying audio narration. By integrating cutting-edge AI technologies, including image captioning (BLIP), story generation (Falcon-7B), and text-to-speech (ESPNet), the

system provides a seamless and interactive multimedia experience. The combination of visual, textual, and auditory elements creates a holistic storytelling platform that is both creative and engaging, catering to various applications in entertainment, education, and accessibility. The intuitive user interface, backed by a robust back-end and scalable cloud integration, ensures smooth processing and real-time interaction.

This project bridges the gap between visual content and narrative generation, offering a unique tool for users to explore stories derived from images. Its potential to serve as an inspiration generator for writers, a learning tool for students, and an inclusive resource for visually impaired individuals highlights its wide-ranging impact. By advancing the integration of AI models and making them accessible through a user-friendly platform, the project contributes to the evolution of interactive and accessible storytelling in the digital age.

REFERENCES

[1] L. M. Botelho, "A guided journey through non-interactive automatic story generation," in Proceedings of the International Conference on Artificial Intelligence and Narrative Systems (ICAINS), 2024.

[2] H. Lovenia, B. Wilie, R. Barraud, S. Cahyawijaya, W Chung, and P Fung, "Every picture tells a story Image-grounded controllable stylistic story generation," in Proceedings of the Conference on Artificial Intelligence and Multimodal Systems (AIMS), 2024

[3] A. Farhadi, M. Hejrati, M. A. Sadeghi, P Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every Picture Tells a Story: Generating Sentences from Images," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024

[4] S. Fotedar, K. Vannisselroij, S. Khalil, and B. Ploeg, "Storytelling Al A Generative Approach to Story Narration," in Proceedings of the European Conference on Artificial Intelligence and Creative Systems (ECAICS), 2024.

[5] M. Theune, N. Slabbers, and F. Helkema, "The automatic generation of narratives," in Proceedings of the International Workshop on Computational Storytelling (IWCS), 2024