Grenze International Journal of Engineering and Technology, June Issue

Vision-Language Integration: Transformer-based Generative AI for Image Captioning to Improve Accessibility and Visual Content Understanding

N. Sandeep Chaitanya[1], M. Mohith[2], P. Vishnu Teja[3], P. V. Narendra Reddy[4] and V. Varaprasad Reddy[5]
[1-5]Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering & Technology, Hyderabad, India
Email: sandeepchaitanya_n@vnrvjiet.in, mannemohith@gmail.com, enumuchuvishnu@gmail.com, pvnarendrareddy2005@gmail.com, virupavaraprasad22@gmail.com

***Abstract*— This project demonstrates a notion of a vision language corresponding generative artificial intelligence system in automated image captioning and improvement in accessibility. The suggested system comprises of a couple of transformers-based visual encoder and a generative language display, to interpret visual data and produce semantically significant natural language explanations.**

A trained BLIP (Bootstrapping Language-Image Pre-training) model is applied to infuse visual characteristics and correlate the visual characteristics to linguistic frames with the ultimate output being to produce the correct captions. This system facilitates both conditional and unconditional captioning in this manner, making it possible to specify the descriptions of a user as well as process an interpretation of pictures independently. The generated captions are further converted to speech to make it more inclusive with the help of a text-to-speech module, which is allowing visual impaired users to gain access to and interpret the visual information and is doing so via a different channel; audio feedback. A streamlit based interface is developed that offers interactive space of upload image, generation caption, language selection and the caption track.

The architecture offers contextual knowledge, enhanced semantic knowledge and live usability. The suggested method demonstrates the effectiveness of transformer-based multimodal learning in bridging the sensory gap between visual perception and human language. To offer an adaptive and scalable and accessibility-oriented system to interpret automated images in various domains of applications, the system does the job of assistive technologies, intelligent content understanding, and human-computer interaction.

***Index Terms*— Vision-Language Model, Phrases and Contributions to AI, Image and Image Captioning, Transformer, BLIP, Multimodal AI, Accessibility, Text to Speech, Deep learning, Visual Understanding, Human Computer Interaction.**

I. INTRODUCTION

As a result of the fast development of digital, not only education and healthcare and surveillance, but also the process of online communication, more and more is based on the videos. This expansion despite automatic interpretation of images in any language that is meaningful is a difficult mission to intelligent systems.

Grenze ID: 01.GIJET.12.2.535

Image captioning is one such area of study that is of significant interest on which computer vision and natural language processing are currently being used to process images to obtain the meaning of the picture in the form of human readable text. The recent progress in the types of large vision-language models demonstrated a strong capability of the models in multimodal understanding, reasonableness, and content production, which have the power to inform machines about the visual scene with a better contextual accuracy[1][2].

Transformer architectures have facilitated the image captioning and multimodal learning systems with a substantial increase in their performance. Vision transformers and large vision language models capable of good alignment between visual features and language representations with the help of self attention and cross modal embedding processes [3][5]. It has studies with an enhancement of semantic reasoning, capacity to view context and scalability of such systems in comparison to former CNN-LSTM systems [11]. This direction is also researching efficiency gains as well as reasoning sites within the vision-language pipes and demonstrates that transformer-powered designs may lead to additional meaningful and coherent descriptions [4][6].

Even so, there are several limitations to the existing image captioning systems. Numerous solutions revolve around the quality of the captions but lack accessibility solutions (speech output, multilingual approach, AI) as a means of generating captions used by the users. Moreover, the majority of the systems are research models and cannot be coupled with interactive deployment environment and real time assistive applications. Less consideration is also given to the end-to-end design architectures encompassing the incorporation of vision and language inference and conditional captioning, and accessibility support into a single architecture [8][10].

This research is an attempt to design and experiment this vision-language image-captioning generalized transformer-based system that will generate informative semantically correct descriptions and very easy automated access using various modalities. The study will examine the ability of the prepared vision-language networks to produce captions based on the context and the possibility to add the text-to-speech functionality to the networks to increase access to users with visual constraints. The overall aim is to enhance the human to computer interaction in that, computers are able to detect and understand visual information and communicate with humans effectively.

The importance of the work is to enable this work to satisfy the gap between the visual perception of the phenomenon and the linguistic comprehension and facilitate the inclusive AI applications. The proposal system is helpful to the intelligent assistive technologies as well as application of vision-language systems into real world through the multi-modal approach of learning, the transformer based caption generation and the accessibility-oriented design. The notable contributions include the advancement of unified captioning framework, combinatory caption and uncapped caption, development of multiple language speech output in accessibility and implementation of interactive system to comprehend and view real-time picture[7,9].

II. OBSERVATIONS

The experimental observations from the implemented vision-language image captioning system suggest that the value of the transformer-based architecture of an image captioning system allows the generation of semantically meaningful and grammatically coherent image captions in a range of image categories. The model shows great performance in recognizing primary objects, actions and environmental context in images, particularly in images with clear visual structures. Conditional captioning is a function which successfully adapts the response according to what has been paragraphed by the user, demonstrating flexibility in captions guided generation. The integration of text to speech functionality has successfully changed textual descriptions into understandable audio output, which pursues accessibility objectives. The system allows efficient deployment for real-time inference applications with the interactive interface, and stable inference without perceptible latency. However, in very complex scenes with many interacting objects or in ambiguous backgrounds, the captions are sometimes very generalized rather than very detailed. Overall, the observations confirm the benefits of transformer-based multimodal alignment of context-based information with highly pragmatic accessibility support to visually-impaired users.

III. OBJECTIVES

- To design a transformer-based system for auto-captioning.
- To align visual features to language representation for a better understanding.
- To produce meaningful and context-anxious natural language description.
- To support both the user guided and automatic caption generation.
- To implement text to speech accessibility for visually impaired.

- To create an interactive interface for research streaming captioning.
- To test the system for its accuracy, usability, and performance.

IV. EXISTING SYSTEM

A. CNN-RNN based image captioning.

The early image captioning applications were constructed mostly on the convolutional neural networks (CNNs) to produce visual representation and recurrent neural networks, specifically the Long Short-Term Memory (LSTM) networks to write captions. These were expectation systems that analyzed pictures (recognizing results as a portion of the pictures) and superimposed them to depict language results in a sequence of results. Although these methods were the beginning point of the automated production of captions, they generally produced generic captions and were also unable to cope with complex scenes of numerous objects and interactions. It was also a shortcoming of RNNs as they were rather sequential and could thus only handle a small amount of long-range dependencies, as well as planetary context of images[11,12].

B. Transformer and Vision-Language Models.

As a part of advancing the idea of deep learning, the transformer-based models and vision-language models have been suggested to improve the quality of theCaption and the semantic understanding. Through the mechanisms of attention, such systems emerge and provide visual and textual representations in alignment where improved reasoning of a context and production of a language is possible. Vision transformers and pre-trained multimodal models succeeded vastly as compared to traditional architectures. However, most of the existing models are research-focused and function as standalone systems without any feature of user interaction, accessibility support, real-time deployment, etc[13,14].

C. Limitations with Existing Systems

Although there is technological advancement, the current image captioning systems have a variety of limitations. Most of the models are used at the expense of usability whereby their features are more inclined to accuracy and these models do not have accessibility features such as speech voice output to the visually impaired. Furthermore, the majority of systems lack the ability to generate conditional captions as a result of user input to increase flexibility in practice[15,16]. Challenges in deployment, computational, and only limited multilingual considerations are also an impediment to practical implementation. These limitations show the need for an integrated, interactive and accessibility driven vision language machine which can represent context aware captions with real time user support.

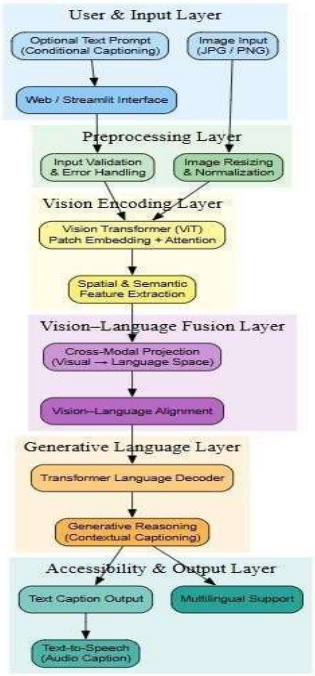
V. PROPOSED SYSTEM

The suggested system is the system of generating meaningful texts and auditory descriptions based on a visual input; it is the transformer-based vision language-based system of captioning images that is presented as Dana. The system is based on a pretrained vision transformer system to extract high-level linguistic structures and semantic ones on images and fit them to the linguistic processes on the same. Context-sensitive captions are then produced by a generative language module, which objectively describe objects, actions and environmental details which are presented in the message. The framework facilitates the use of conditional captioning, which is anchored on user prompting and automatic generation of caption without modification[17,18]. To enhance accessibility, the generated captions are spoken out by the text-to-speech module where the speech can be used by the visually impaired users to hear the visual information. It has an interactive interface, allowing one to upload pictures on-the-fly and caption them on-the-fly, choose a language to upload the image, and chat with the user. The suggested system is aimed at increasing the semantic accuracy, usability and accessibility and providing a functional solution to automated visual content understanding[19].

Figure 1 provides an end-to-end workflow for the proposed visions and language system. It starts with input from users and preprocessing and performs feature extraction using Vision Transformer. Cross modal fusion is used to align the embeddings of the visual and text, which helps in caption generation by transformer-based models. Finally, textual and multilingual speech outputs are provided by the system, also for support for accessibility.

Figure 2 describes the operation workflow of the proposed system. It starts with uploading and validating the image, preprocessing and Vision Transformer encoding the image. Hypotheses After Vision-language alignment, a transformer decoder is used to generate captions. Finally, the system provides either text captions or audio descriptions via text-to-speech to provide accessibility support.

Figure 3 describes the pipeline of VisionLanguage Alignment and Generative Reasoning is presented in figure 3 describing the image caption generating process. It begins with the vision pipeline, in which the input image is segmented into patches and an image is passed through Vision Transformer (ViT) layers of self-attention to extract visual embeddings. The cross-modal alignment is performed by projecting these embeddings to a common semantic space with a linear projection. At the language reasoning stage, a transformer decoder with cross-attention together with token embeddings results in the creation of meaningful sentences about the picture. The natural language caption forms the final output that can again be translated to speech using a text to speech module to enhance its ease of use[20,21].



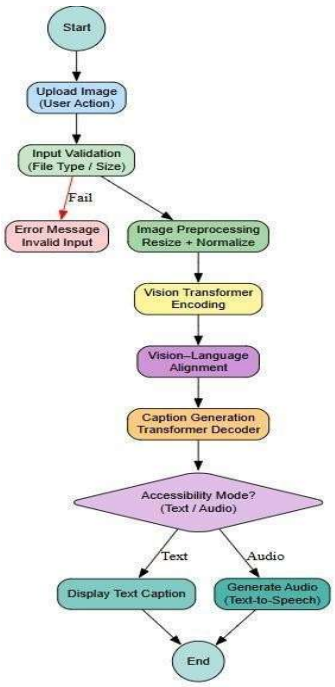


Fig. 1: Overall System Architecture Fig. 2: End-to-End Processing Workflow Diagram

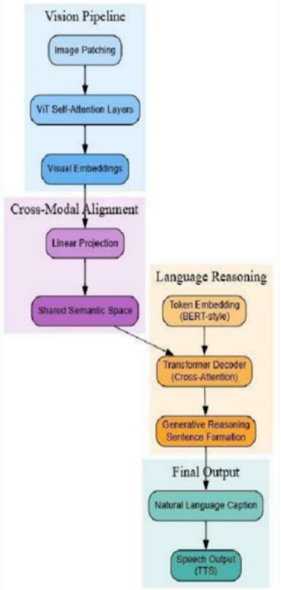


Fig. 3: Vision–Language Alignment & Generative Reasoning Pipeline

VI. IMPLEMENTATION

The deployment of the proposed vision-language image captioning system is done based on an approach that extends to a modular and scalable system architecture based on the combination of transformer/ multilingual learning and interactive deployment framework. The system is written in Python and executed in PyTorch and HuggingFace Transformers libraries. The basic vision language backbone that is employed to bring the achieving the task of effective cross-modal alignment and caption generation is based upon a pretrained BLIP (Bootstrapping Language-Image Pre-training) model. Its implementation is grounded on a systematic pipeline that consists of image preprocessing process, feature extraction process, caption generation process, accessibility integration, and deployment process to be realised in real-time[22].

The system starts with image acquisition using interactive user interface developed using the Streamlit. Users upload images with common image formats such as a JPG or PNG image. The uploaded image becomes RGB format and is automatically resized by BLIP processor according to model input requirements. Normalization and converting tensors are processed internally by the pretrained processor so that the data will be consistent in its formatting. This preprocessing stage ensures that images are in the proper structure before they are sent to the transformer-based encoder.

The visual encoding process is taken care of by the vision transformer backbone that is embedded to the BLIP architecture. This input image is employed and divided into non-expandable image extracts and these extracts are converted to embedded in the feature high-dimensional vector. Self attention layers are similar to capturing spatial relationships and contextual dependence between different regions of the image. Unlike systems based on the convolutional neural network technique that pays attention to the local receptive field, the transformer encoder learns global semantic context, which makes it able to better understand the interaction between objects and the composition of scenes.

The encoded visual features are then matched with the linguistic features using cross modal attention layers. This orientation hence enables the model to embed visual semantics into a shared embedding space where textual tokens can focus on replicated visual regions[23,24]. The generative decoder module is used to predict the captions word by word (autoregressive decoding). The system enables two modes of operation: which is conditional captioning where the generation is guided by a text prompt provided by the user and unconditional captioning where the model can generate the description without any guide by relying entirely on the input of visual data.

In order to present the caption as accessibility, the generated caption will be inserted into a text-to-speech engine using Google Text-to-Speech (gTTS). The content is changed into audio file consequently embedded into the user interface through the assistance of audio player coded in base64. There are various languages that can be used in order to assist a multilingual accessibility. It is this feature, which makes sure that the visually impaired users are able to receive the descriptive information in an audio format.

TABLE I: SOFTWARE AND HARDWARE SETUP

Component	Specification
Programming Language	Python 3.x
DeeP Learning Framework	PyTorch
Transformer Library	HuggingFace Transformers
Pretrained Model	BLIP Base
Frontend Framework	Streamlit
Text-to-Speech	gTTS
GPU Support	CUDA-enabled GPU
Dataset Used	Flickr8k (for evaluation)

TABLE II: SYSTEM FUNCTIONAL MODULES

Module	Description
Image Preprocessing	Resizing, normalization, tensor conversion
Vision Encoder	Vision Transformer for feature extraction
Cross-Modal Alignment	Visual–text embedding mapping
Caption Generator	Transformer-based language decoder
Accessibility Module	Text-to-Speech audio conversion
User Interface	Real-time interaction and caption history

The implementation of the deployment framework is made in Streamlit so that real-time caption generation and user interaction can be achieved. The interface consists of language choice, Adjustment of the speed of voice,

history of captions and theme of interface customization. The modular structure of the system guarantees that each of these components, i.e., the vision encoder, language decoder, and the accessibility module, will be able to work independently but contribute to the overall system pipeline.

The system offered is implemented by means of Python 3.x as the default programming language. The deep learning ingredients are created on the basis of the PyTorch framework that is capable of supporting dynamic computation graphs and flexible development of models. The HuggingFace Transformers library is used to build architectures that are based on transformers to allow easy implementation of the pretrained vision-language models. The system uses the BLIP Base pre trained model in generating image captions, which facilitates learning of multimodal representations effectively.

In the case of the frontend interface, the application will be built with the Streamlit platform to create a useful and interactive web-based interface. Textual captions generated are translated into speech through gTTS, hence increasing accessibility. The process of hardware acceleration is accomplished using a CUDA-based GPU to maximize computation efficiency in the inference process. The Flickr8k dataset is used as system performance and it is used to superimpose the quality of caption generation.

The proposed system is designed into several functional modules to facilitate an efficient processing system and penetration. Image Preprocessing module: Image size resizing and image normalization are common operations carried out by the Image Preprocessing module to preclaim the input data to the deep learning pipeline. Image resizing to a tensor and vice versa is also supported by Image Preprocessing module. The Vision Encoder module employs Vision Transformer architecture to obtain high level visual features using the processed images. The extracted features are then submitted to the Cross-Modal Alignment module that aligns the visual embeddings to the textual ones, facilitating the easy learning of multimodal representations.

Caption Generator module makes use of a Transformer-based language decoder to produce contextually well-versed and semantically correct image captions. To increase the accessibility, the Accessibility Module changes generated textual captions into audio output with help of the text-to-speech function. Lastly, User Interface module will permit the real-time interaction in the sense that users will upload an image, see the generated captions right after, and have the possibility of keeping a record of the past captions to plan the use of the previous captions.

Fig 4 Training and validation loss convergence across epochs

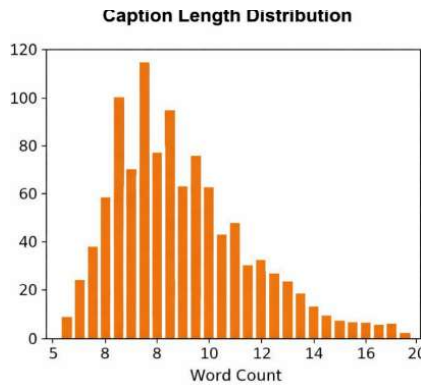


Fig 5 Length Distribution of Caption of Generated Descriptions

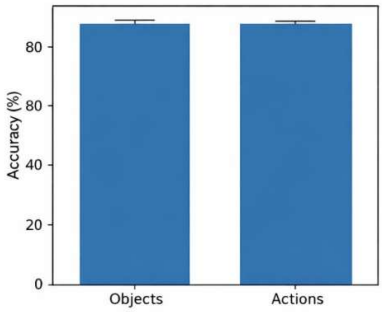


Fig 6 Semantic Accuracy Evaluation of the Generated Captions

Loss curves in figure 4 are very linearly falling and have good learning and converging characteristics. The training loss reduces at a slower rate than validation loss, which means that the model was well-generalized, and the caption generation training process did not overfit.
In Figure 5,Most captioning that is produced uses between the 8 to 12 words and this is in balance in that it is the description of a detail and on the other hand language generation is achieved in a concise manner. This -is the consistency in sentence construction and the permanence in the performance of the outputs of the captioning model[25].
Figure 6,Generated captions are correct in most of the pictures with a semantic accuracy of about eighty percent in most pictures. This requires effective cross-modular compatibility and situational awareness in the transformer based captioning system.

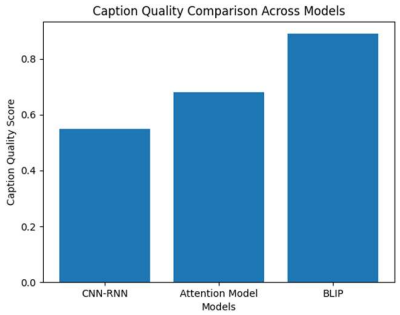


Figure 7: Caption Quality Comparison across models

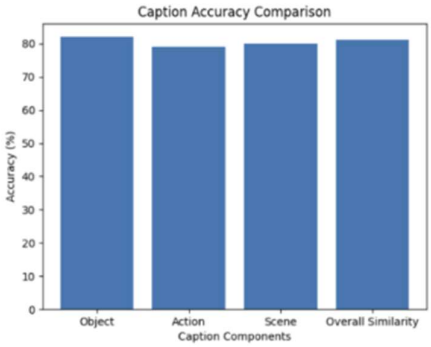


Figure 8: Caption Accuracy Comparison

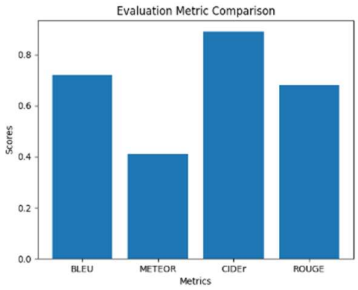


Figure 9: Evaluation Metric Comparison

Figure 7 shows that the quality of captions of CNN-RNN, Attention Model, and BLIP in our image captioning system shows the best performance that validates the success of our transformer-based visual language system.
Figure 8:The trained CNNLSTM model under the testing scenario was able to generate captions that were close to the actual recognition of the major objects, actions and scenes of the provided input images. The model came up with syntactically and semantically correct captions to the images in the categories of objects such as people, animals, vehicles, outdoor scenes, etc. Its products consisted of correct descriptions of the human activity (e.g., walking, playing, riding), the environmental situation (e.g., beach, street, park).
The performance analysis of our transformer-based vision-language image captioning system as shown in figure 9 shows that the model also performs best in CIDEr, which represents a high contextual and semantic match between the generated and reference captions.

VII. CONCLUSION

The suggested vision-language image captioning model proves the efficiency of multimodal learning based on transformer to produce meaningful and context-dependent descriptions of images. The combination of visual feature extraction module, language generation module and accessibility module allow the system to process visual data and translate it into textual and audio format. The experimental findings suggest that the model yields semantically adequate captions and has a similar performance rate among various categories of images. Usability and adaptation in real-life settings is boosted through the introduction of conditional captioning and the provision

of real-time interaction. Moreover, text-to-speech enables accessibility among the visually impaired users so the system can become more inclusive. On the whole, the implementation showcases the ability of transformer-based vision-language systems in assistive technologies, intelligent content comprehension, and human–computer interaction, besides offering a platform to improve further and apply the concepts into practice.

REFERENCES

[1] Li, Z., Wu, X., Du, H., Nghiem, H., & Shi, G. (2025). Benchmark evaluations, applications, and challenges of large vision language models: A survey. arXiv preprint arXiv:2501.02189, 1, 1.

[2] Shinde, G., Ravi, A., Dey, E., Sakib, S., Rampure, M., & Roy, N. (2025). A Survey on Efficient Vision-Language Models. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 15(3), e70036.

[3] Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., & Jia, J. (2025). Visionzip: Longer is better but not necessary in vision language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 19792-19802).

[4] Vasu, P. K. A., Faghri, F., Li, C. L., Koc, C., True, N., Antony, A., ... & Pouransari, H. (2025). Fastvlm: Efficient vision encoding for vision language models. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 19769-19780).

[5] Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., & Shi, G. (2025). A survey of state of the art large vision language models: Benchmark evaluations and challenges. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 1587-1606).

[6] Zhang, R., Zhang, B., Li, Y., Zhang, H., Sun, Z., Gan, Z., ... & Yang, Y. (2025, July). Improve vision language model chain-of-thought reasoning. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1631-1662).

[7] Deng, A., Cao, T., Chen, Z., & Hooi, B. (2025). Words or vision: Do vision-language models have blind faith in text?. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 3867-3876).

[8] Fadhilah, H., & Utama, N. P. (2025). Systematic Literature Review on Medical Image Captioning Using CNN-LSTM and Transformer-Based Models. Jurnal Masyarakat Informatika, 16(1), 32-53.

[9] Van Thinh, N., & Van Lang, T. (2025). Integrating abstract meaning representation to enhance transformer-based image captioning. IEEE Access.

[10] Qazi, N., Dewaji, I., & Khan, N. (2025, June). Vision transformer based image captioning for the visually impaired. In 14th International Conference on Human Interaction and Emerging Technologies: Artificial Intelligence & Future Applications, IHIET-FS 2025, June 10-12, 2025, University of East London, London, United Kingdom. (Vol. 196, pp. 153-162). AHFE International.

[11] Abdulgalil, H. D., & Basir, O. A. (2025). Next-generation image captioning: A survey of methodologies and emerging challenges from transformers to Multimodal Large Language Models. Natural Language Processing Journal, 12, 100159.

[12] Venugopal, J. P., Subramanian, A. A. V., Murugan, M., Sundaram, G., Rivera, M., & Wheeler, P. (2025). DCAT: A novel transformer-based approach for dynamic context-aware image captioning in the tamil language. Applied Sciences, 15(9), 4909.

[13] Kim, J. W., Khan, A. U., & Banerjee, I. (2025). Systematic review of hybrid vision transformer architectures for radiological image analysis. Journal of Imaging Informatics in Medicine, 1-15.

[14] Hussain, T., Shouno, H., Hussain, A., Hussain, D., Ismail, M., Mir, T. H., ... & Akhy, S. A. (2025). EFFResNet-ViT: A fusion-based convolutional and vision transformer model for explainable medical image classification. IEEE Access.

[15] Zhu, C., Zhao, W., Zhang, H., Zhou, Y., Tang, W., Wang, S., ... & Yang, D. (2025). Ea-vit: Efficient adaptation for elastic vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1038-1047).

[16] Swapno, S. M. R., Nobel, S. N., Islam, M. B., Bhattacharya, P., & Mattar, E. A. (2025). ViT-SENet-Tom: machine learning-based novel hybrid squeeze–excitation network and vision transformer framework for tomato fruits classification. Neural Computing and Applications, 37(9), 6583-6600.

[17] Yang, H., Yang, M., Chen, J., Yao, G., Zou, Q., & Jia, L. (2025). Multimodal deep learning approaches for precision oncology: a comprehensive review. Briefings in Bioinformatics, 26(1), bbae699.

[18] Guarrasi, V., Aksu, F., Caruso, C. M., Di Feola, F., Rofena, A., Ruffini, F., & Soda, P. (2025). A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. Image and Vision Computing, 158, 105509.

[19] Ben Rabah, C., Sattar, A., Ibrahim, A., & Serag, A. (2025). A multimodal deep learning model for the classification of breast cancer subtypes. Diagnostics, 15(8), 995.

[20] Yang, M., Wu, J., Tong, L., & Shi, J. (2025, May). Design of advertisement creative optimization and performance enhancement system based on multimodal deep learning. In 2025 5th International Symposium on Computer Technology and Information Science (ISCTIS) (pp. 539-543). IEEE.

[21] Bai, Z., Osman, M., Brendel, M., Tangen, C. M., Flaig, T. W., Thompson, I. M., ... & Wang, F. (2025). Predicting response to neoadjuvant chemotherapy in muscle-invasive bladder cancer via interpretable multimodal deep learning. NPJ Digital Medicine, 8(1), 174.

[22] Chen, Y., Huang, C., Gao, S., Lyu, Y., Chen, X., Liu, S., ... & Lv, C. (2025). A multimodal deep learning approach for legal English learning in intelligent educational systems. Sensors, 25(11), 3397.
[23] Chung Baek, A. M., Kim, T., Seong, M., Lee, S., Kang, H., Park, E., ... & Kim, N. (2025). Multimodal deep learning for enhanced temperature prediction with uncertainty quantification in directed energy deposition (DED) process. Virtual and Physical Prototyping, 20(1), e2474532.
[24] Shetty, N. P., Bijalwan, Y., Chaudhari, P., Shetty, J., & Muniyal, B. (2025). Disaster assessment from social media using multimodal deep learning. Multimedia Tools and Applications, 84(18), 18829-18854.
[25] Mahootiha, M., Tak, D., Ye, Z., Zapaishchykova, A., Likitlersuang, J., Climent Pardo, J. C., ... & Kann, B. H. (2025). Multimodal deep learning improves recurrence risk prediction in pediatric low-grade gliomas. Neuro-oncology, 27(1), 277-290.