Transformers and LLMs as the New Benchmark in Early Cancer Detection

Yulia Kumar[1], *Kuan* Huang[1], *Zachary* Gordon[1], *Lais* Castro[1], *Egan* Okumu[1], *Patricia* Morreale[1] and *J. Jenny* Li[1]

[1]Department of Computer Science and Technology, Kean University, Union NJ 07083, USA

Abstract. The study explores the transformative capabilities of Transformers and Large Language Models (LLMs) in the early detection of Acute Lymphoblastic Leukaemia (ALL). The researchers benchmark Vision Transformers with Deformable Attention (DAT) and Hierarchical Vision Transformers (Swin) against established Convolutional Neural Networks (CNNs) like ResNet-50 and VGG-16. The findings reveal that transformer models exhibit remarkable accuracy in identifying ALL from original images, demonstrating efficiency in image analysis without necessitating labour-intensive segmentation. A thorough bias analysis is conducted to ensure the robustness and fairness of the models. The promising performance of the transformer models indicates a trajectory towards surpassing CNNs in cancer detection, setting new standards for accuracy. In addition, the study explores the capabilities of LLMs in revolutionising early cancer detection and providing comprehensive support to ALL patients. These models assist in symptom analysis, offer preliminary assessments, and guide individuals seeking information, contributing to a more accessible and informed healthcare journey. The integration of these advanced AI technologies holds the potential to enhance early detection, improve patient outcomes, and reduce healthcare disparities, marking a significant advancement in the fight against ALL.

1 Introduction

Acute Lymphoblastic Leukaemia (*ALL*) predominantly affects individuals under 20, necessitating early detection to combat its aggressive nature. Current diagnostic methods are invasive, costly, and time-consuming, with a significant emotional impact on patients and families. In 2023, the American Cancer Society estimates 6,540 new *ALL* cases and 1,390 deaths in the U.S., highlighting the urgency for innovative solutions [1]. ALL risk factors vary with age, peaking in young children and rising again after 50 [2]. Despite constituting less than 0.5% of all U.S. cancers, most ALL-related deaths occur in adults, underscoring the need for improved diagnostic methods [1, 3]. Artificial Intelligence (AI) is revolutionizing this field, with machine learning (ML) models like Vision Transformers with Deformable Attention (DAT), Hierarchical Vision Transformers (Swin), and Large Language Models (LLMs) like ChatGPT, Bing and Bard showing promising results. This study aims to set new standards in accuracy, affordability, and speed of ALL diagnoses, addressing three main research questions: *RQ1: How do transformer models compare to CNNs in early cancer detection? RQ2: What role can LLMs play in ALL detection? RQ3: How can biases in AI models trained on ALL data be identified and mitigated?* The research aspires to enhance patient well-being and contribute to societal health advancements.

2 Related Work

The adoption of Transformers and Large Language Models (LLMs) in healthcare, especially for early cancer detection, including Acute Lymphoblastic Leukaemia (ALL), represents a significant advancement. This synergy has catalysed research, diagnosis, and treatment innovations, supported by an expanding literature base. Nerella et al. provided a thorough survey on Transformers in healthcare, highlighting their versatility across fields like genomics and patient care [4]. Balabin introduced Multimodal Transformers, showcasing their capability to manage complex biomedical data [5]. Thirunavukarasu et al. discussed LLMs in medicine, emphasizing the necessity for models that comprehend complex medical terminology [6]. Holmes et al. demonstrated LLMs' effectiveness in niche areas like radiation oncology physics [7]. Wang et al. and Li et al. explored ChatGPT's integration into biomedical research and healthcare, indicating LLMs' increasing acceptance in medicine [8-9]. Batzoglou focused on LLMs in Molecular Biology, illustrating their potential to unravel complex biological language and contribute to understanding genetic factors in diseases like cancer [10]. In early cancer detection, Transformers have been central. The Kaggle ALL dataset authors-researchers highlighted DenseNet-201's performance with various CNN architectures [11]. Newer transformers like DAT [12] and Swin [13], and LLMs like ChatGPT-4, have expanded early ALL detection possibilities. Huang et al.'s work on breast cancer using transformers adds valuable insights into early detection methods [14-15]. Addressing bias in Transformers, crucial across neural networks, has been

intensively tackled with models like DETR and Deformable DETR [16-18]. Kumar et al.'s testFAILS framework guides responsible LLM apps [19].

3 Methodology

3.1 The project dataset

The ALL dataset, pivotal for this study, comprises a well-curated collection of medical images, transformed and annotated for training AI/ML models. Sourced from Kaggle [11], it underpins our exploration of AI in early Acute Lymphoblastic Leukaemia (ALL) detection and classification. The dataset encompasses 3,256 Peripheral Blood Smear (PBS) images from 89 individuals, 25 healthy and 64 with varying ALL stages, prepared by experienced lab personnel. It categorizes images into benign (non-progressing cancer) and malignant classes, the latter further divided into Early Pre-B, Pre-B, and Pro-B ALL subtypes, representing different ALL stages. Notably, the dataset adheres to a 1970s classification system (L1, L2, L3), now outdated and misaligned with current American Cancer Society and WHO guidelines, but still in use at the dataset's originating laboratory outside the US. The images, captured via microscopy, underwent colour thresholding-based segmentation to identify cell types and subtypes. However, this method showed limitations, struggling with accurate blast cell region definition, background noise, and segmentation in ALL images [11]. Table 1 provides an overview of the dataset. Figure 1 displays sample images and their corresponding segmented versions.

Table 1. The project dataset.

Types of ALL	The # of images	Comments
Benign	504	Noncancerous
Early	985	Early-stage type (L1)
Pre	963	Middle stage type (L2)
Pro	804	Later stage type (L3)
Classification	4 classes	Benign, Early, Pre, Pro

As can be seen from Table 1, the Benign ALL type is the least dangerous as it is not spreading to the neighbouring cells, and in many online resources, it is called noncancerous, while the other three are cancerous and present different types/stages of ALL. The examples of ALL images can be seen below:

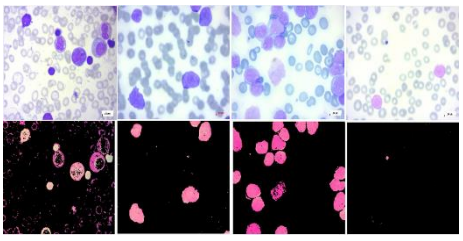

Fig. 1. The project ALL dataset (top: original benign, early, pre, pro images; bottom: corresponding segmented images).

The segmented images revealed the need for sophisticated AI models capable of classifying original, non-segmented ALL images, marking a departure from previous dataset applications. The ALL dataset [11] thus emerges as a valuable resource for advancing AI in ALL diagnosis and classification, highlighting the necessity for innovative, accurate AI techniques to improve early cancer detection.

3.2 Applying CNNs and Transformers to ALL Data

In this study, the traditional convolutional neural networks such as VGG-16 and ResNet-50 and advanced transformer models, specifically DAT [12] and Swin Transformers [13], were employed on the ALL dataset to evaluate their performance in achieving optimal accuracy. The hypothesis posited that both transformer models would be highly compatible with the dataset and outperform the convolutional neural networks. The dataset's unique features, such as image presentation, segmentation challenges, and oval cell shapes, were deemed favourable for computer vision tasks. Training the DAT transformer on computer vision tasks, particularly the ALL dataset, was a complex endeavour. After the initial 40 epochs, the model achieved a 93.56% accuracy on the test set. Figure 2 illustrates the losses:

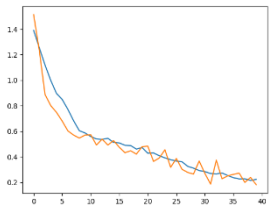

Fig. 2. Initial Training and Validation Losses of the DAT

For comparative purposes, the researchers ran the convolutional models (VGG-16 and ResNet-50) alongside the transformers. Table 2 presents the results:

Table 2. Comparative Results of various models.

Model	Accuracy, %				
	Benign	Early	Pre	Pro	Average
ResNet-50	89.11	83.25	98.45	100	92.7025
VGG-16	99.01	99.49	100	100	99.6250
DAT	89.11	96.95	99.48	100	96.3850
Swin	92.08	99.49	99.48	100	97.7625

As can be seen from the table DAT and Swin outperformed ResNet-50 on initial images across all classes, particularly in early cancer detection, where they achieved 97-99% accuracy compared to ResNet-50's 83%. Through VGG-16 model performed on early cancer detection better. This underscores the new generation models' perspectives to bypass the costly and time-intensive segmentation step for ALL data. The necessity of the fine-tuning is evident. The experimental parameters are detailed in Table 3:

The project ALL dataset (top: original benign, early, pre, pro images; bottom: corresponding segmented images).

Table 3. Experimental Parameters.

Parameter	Comment
Initial Dataset	3256 images
Class	80% for training, 20% for testing at random
Batch Size	4 classes (Benign, Early, Pre, Pro)
Input Image	16
Optimizer	256×256 (resized)
# of Epochs	SGD, learning rate 0.001.
PyTorch	300
Hardware	1.12.1.
	Ubuntu 20.04.5 Linux system : AMD EPYC 7513 32-Core Processor 2.60GHz, 8 NVIDIA GeForce 3090 graphics cards, each one of 24 Gb

The Swin and DAT Transformers achieved accuracies of 97.76% and 96.385%, respectively, outperforming ResNet-50 but also highlighting the challenges in achieving perfect accuracy in early ALL stages, even with cutting-edge transformers. Confusion matrices were generated to further analyse performance:

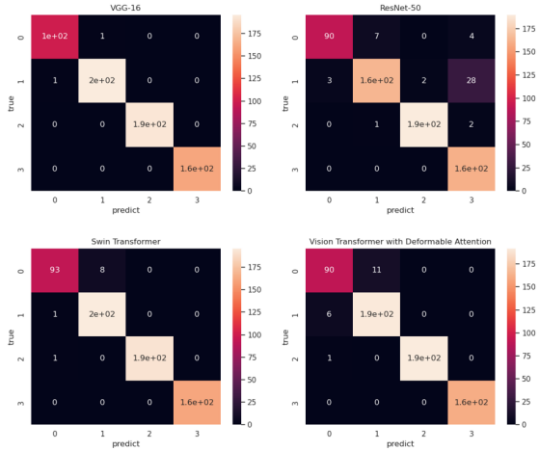

Fig. 3. Confusion Matrices for VGG-16 (top left), ResNet-50 (top right), Swin (left bottom) and DAT (right bottom).

As can be seen from Figure 3, the Swin Transformer outperformed the DAT, contradicting the original DAT paper [12] and highlighting the attention architectures' sensitivity to specific datasets and tasks. Training and testing loss curves for DAT, Swin, and ResNet-50 provide additional insights (see Figure 4):

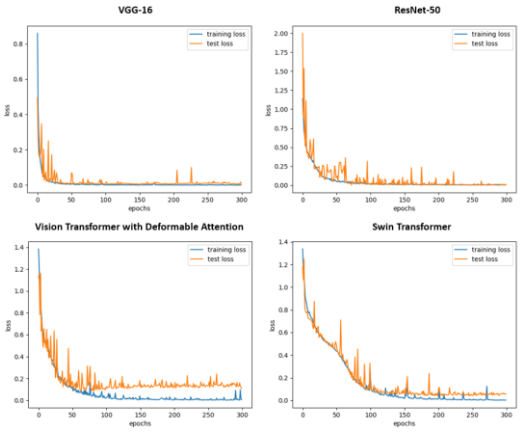

Fig. 4. Loss Curves for VGG-16, ResNet-50, DAT, and Swin.

The results contribute to our understanding of ALL detection and pave the way for future research and innovation in applying transformer models to medical imaging and diagnosis.

3.3 The role of LLMs in early cancer detection

The exploration of Large Language Models (LLMs) in the detection of Acute Lymphoblastic Leukaemia (ALL) is a burgeoning field, showing great potential despite its experimental nature. An attempt to utilize ChatGPT-4's code interpreter for ALL detection in cell images did not yield direct results in terms of classification. However, the model demonstrated its utility by providing a comprehensive metadata analysis of the image and generating executable code for a pre-trained Faster R-CNN model, which was successfully run in Google Collab [20]. The multimodal capabilities of LLMs are rapidly evolving, with anticipation building around the upcoming ChatGPT-5 model, which is expected to handle a variety of media data types, including video and 3D data. Image analysis and generation capabilities are already present in models like DALL-E 3 and ChatGPT-4, showcasing the versatility of these technologies.

LLMs have proven their worth as personal assistants, offering invaluable support to ALL patients in numerous ways. A notable development in this area is the AssureAIDoctor (AAID) app, a novel tool designed to enhance patient care [21], recently upgraded to .NET 8 and ChatGPT-4. The prototype's GUI is displayed in Figure 5.

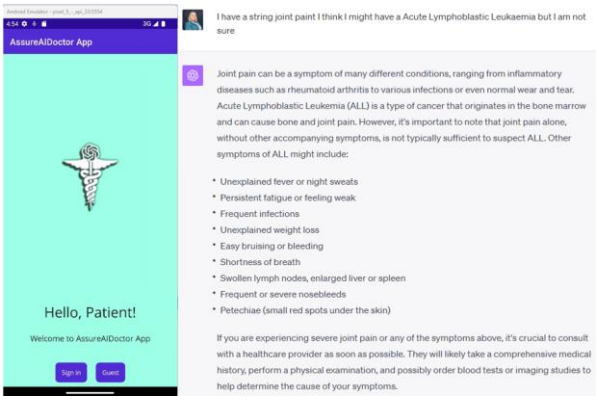

Fig. 5. The AssureAIDoctor's AI assistant for ALL patients.

The app has a potential to help ALL patients in many ways. Some of them are provided in Table 4.

Table 4. Enhancing the Patient Experience with AI Doctor.

Feature	**Description**
Personalized Patient Interaction	Provides tailored conversations based on symptoms and medical history, aiding in early detection and guidance for ALL patients or those suspecting they might have ALL.
Symptom Analysis and Preliminary Diagnosis	Analyses symptoms, compares them with known symptoms of ALL, and provides a preliminary assessment to serve as a quick and accessible first step for individuals seeking information and guidance.
Image Analysis and Assistance	Assists users in understanding their medical images, such as blood smears, through initial

	analysis, helping to demystify complex medical data.
Information and Education	Offers educational content on ALL, empowering patients with knowledge to aid in informed decision-making.
Emotional Support and Mental Health	Provides emotional support, engages in conversations to alleviate anxiety, and guides users to professional mental health resources if necessary.
Facilitation of Doctor-Patient Communication	Helps organize medical information, symptoms, and questions for discussions with healthcare providers, ensuring effective use of medical appointments.
Follow-Up and Medication Reminders	Assists in follow-up care, providing reminders for medication, appointments, and necessary tests to help ALL patients stay on top of their treatment plan.
Community and Support	Connects ALL patients with support communities, offering a platform for sharing experiences, advice, and encouragement.
Research and Data Collection	With user consent, collects anonymized data on ALL symptoms and progression, contributing to a broader understanding of the disease and aiding in research efforts.
Accessibility and Reducing Healthcare Disparities	Ensures individuals from all backgrounds have access to quality healthcare information, providing immediate access to information and preliminary assessments, and thus helping in reducing healthcare disparities.

ChatGPT-4-Turbo, Bing and Bard bots were employed to analyse ALL data directly through their multimodal capacities. Their responses, which include image descriptions and identification of various cell types, demonstrate their ability to extract information without prior training, as shown in Figures 6 and 7.

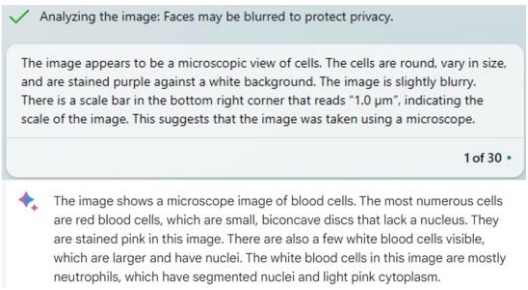

Fig. 6. Bing (top) and Bard (bottom) Responses.

ChatGPT4-Turbo model was fed three images into: one non-cancerous, one with early cancer and one to analyse. Model's response is provided below:

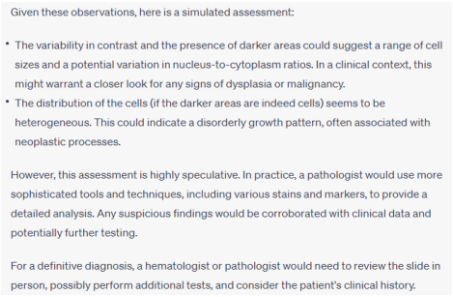

Fig. 7. ChatGPT-4-Turbo response [22].

The ChatGPT-4-Turbo model closely analysed the image and even created its threshold, what can be seen in the recorded conversation [22], but it refused to answer a yes or no question about the presence of early cancer in the anonymous image (that was present there).

The integration of LLMs such as ChatGPT, Bard, and Bing, alongside innovative tools like the AssureAIDoctor, holds great potential in advancing ALL research and patient care. These technologies are becoming increasingly prevalent across various domains, showcasing their versatility and potential to revolutionize healthcare. The app powered by ChatGPT and DALL-E, offers a personalized and comprehensive approach to patient care, from symptom analysis to preliminary diagnosis, and even image analysis. This integration is setting new standards in accuracy, efficiency, and innovation, paving the way for a future where healthcare is more accessible, personalized, and effective for ALL patients. As illustrated above, the application of LLMs to ALL data shows promise, but it necessitates further research and validation to ensure reliability and accuracy.

4 Bias Detection in Transformers

Addressing biases in transformer models, specifically within the Multilayer Perceptron (MLP) of the first transformer stage of both DAT and Swin models, was a critical and complex aspect of this study. This section provides an in-depth analysis and discussion of the extensive trials and evaluations conducted to understand and mitigate these biases. Figure 8 presents a comparative analysis of the biases of the transformer models. The Swin Transformer and DAT models required nearly 100 epochs to reach convergence, highlighting the intricate nature and variability in training such models.

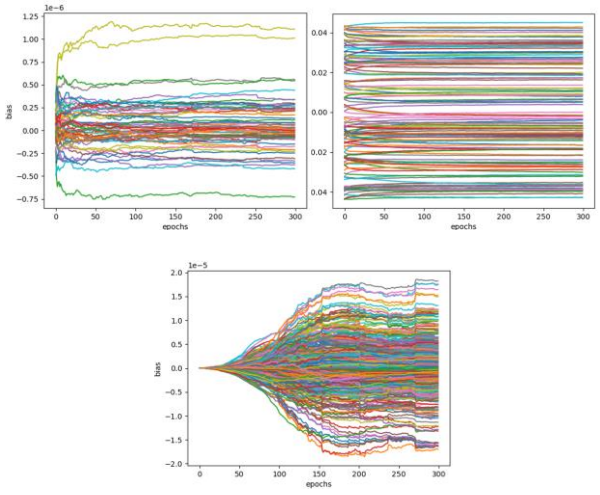

Fig. 8. Bias Visualizations of VGG-16 (the first convolutional layer), and DAT and Swin Models (MLP of Second Transformer Stage).

The biases in the Swin transformer exhibited a unique dome-like shape; however, the VGG-16 and DAT's biases are more stable. The bias of the Swin transformer shows the hardness of training convergence in the Swin transformer. It takes 150 epochs to get stable values. According to Bard bot this is because the model has learned a bias towards dominant features in the

training data. This can be a problem for tasks where it is important to be unbiased, such as for medical diagnosis. According to the bot the model's biases may depend on the training data, meaning that the training data has appeared to be biased. There might be an overfit. The biases in the second MLP of the last transformer stage of the DAT model stabilized around epochs 45-50, forming almost a straight line, indicating a stabilization in the learning process. On the other hand, the Swin Transformer required close to 200 epochs for the biases to stabilize, emphasizing the model's complexity and the nuanced interplay of its internal parameters.

A noteworthy discovery in this research was the behaviour of the ResNet-50 CNN model, which exhibited no biases in its first convolutional layer:

$$self.conv1 = nn.Conv2d(3, self.inplanes, kernel\_size=7, stride=2, padding=3, bias=False) \quad (1)$$

The ResNet-50 CNN model, with its first convolutional layer set to bias=False, demonstrates a commendable ability to learn unbiased features from the input data. This configuration, especially in conjunction with batch normalization, proves advantageous by eliminating redundancy and reducing the model's complexity. The chosen parameters, including a smaller filter size and stride, ensure the capture of finer details, contributing to a more nuanced understanding of the input data. Enough filters further guarantee the model's capacity to discern a diverse array of patterns, aiding in the mitigation of biases. The careful initialization of weights, alongside the application of regularization techniques, solidifies the stability of the learning process, preventing overfitting and ensuring a fair and balanced representation. This finding is significant as it demonstrates that through careful simulations and tuning, biases can be completely mitigated. It also raises important questions about the inherent differences in how CNNs and transformer models handle biases.

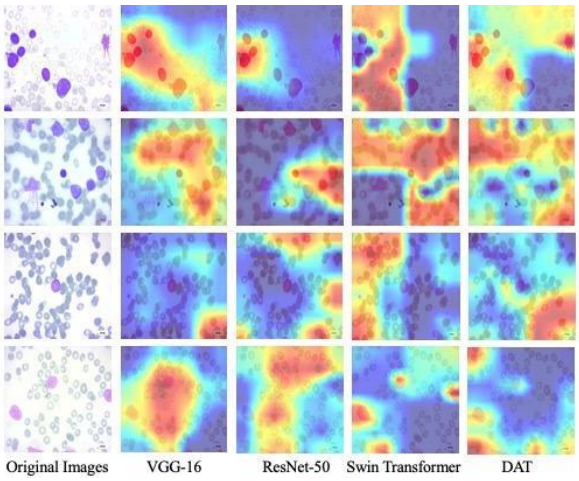

Fig. 9. CAM Results of DL Models.

Class Activation Map (CAM), a technique commonly used for visualizing and interpreting the decisions made by CNNs in computer vision tasks, was employed in this study [23]. CAM generates a heat map highlighting the significant regions of an input image that contribute to a deep Neural Network's classification decision. The researchers utilized the Grad-CAM method to generate CAMs for each classification model in the study, including the convolutional block 5 in VGG-16 and ResNet-50, and the transformer stage 4 in the Swin Transformer and DAT Transformer.

The CAMs, as shown in Figure 9, reveal the regions of importance for making the final classification decisions. The CNN models (VGG-16 and ResNet-50) tended to focus more on specific features in the original images, such as the purple dots, which are significant regions marked by the dataset's creators. In contrast, the transformer models adopted a different approach.

Bard bot was asked to analyse the CAM image and produced the following result:

> It is interesting to note that the images generated by the different models have different color schemes. This suggests that the models may have different biases when it comes to color. For example, the ResNet-50 and Swin Transformer models may be biased towards generating images with bright and saturated colors, while the DAT model may be biased towards generating images with dark and muted colors.

Fig. 10. Bard bot's analysis of CAM Results.

Analysing the CAMs can aid doctors and patients in focusing on the highlighted activations, contributing to the accuracy of diagnosis. It is evident that utilizing transformers in ALL detection can enhance results and provide new perspectives for both doctors and AI specialists. The researchers acknowledge that there is potential for further tuning and optimization of the DAT and Swin models to improve their accuracy and reduce bias. Techniques such as applying various activation functions, adjusting neuron weights, and others are currently under investigation. This ongoing effort is crucial for advancing the field and ensuring the reliability and fairness of AI models in medical apps.

5 Conclusion and Future Work

This comprehensive study has meticulously addressed the outlined research questions, shedding light on the integration of transformer models, Large Language Models (LLMs), and bias mitigation strategies in Acute Lymphoblastic Leukaemia (ALL) detection. The employment of transformer models, specifically DAT and Swin, yielded impressive accuracies of 96.385% and 97.76%, respectively. However, these models required a substantial number of epochs to converge, with Swin necessitating up to 200 epochs for bias stabilization, and DAT showcasing the complexity of training transformer models.

In comparison, the traditional CNN model demonstrated quicker convergence, with significant loss reduction observed between the 10th and 15th epochs. LLMs, represented by ChatGPT-4, Bing, and Bard, exhibited their potential in image classification and AI code generation, with ChatGPT-4 generating a working code for a pre-trained Faster R-CNN model and Bard providing descriptive image analysis. These capabilities underscore the versatility of LLMs in medical imaging, though they necessitate further research and validation for reliability and accuracy assurance.

The study also delved into biases present in the MLP of the first transformer stage of both DAT and Swin models, revealing distinctive dome-like shape in the biases and significant fluctuation in the DAT Transformer models. The identification of the point of convergence and the demonstration that biases could be completely mitigated, as evidenced by the ResNet-50 CNN model, underscore the performance of transformer models, and highlight the promising capabilities of LLMs in medical imaging and explainable AI.

Looking forward, the integration of transformer models and LLMs heralds a future of accurate, humane, accessible, and patient-centric AI in medical diagnostics, contributing to improved healthcare outcomes and fostering a more inclusive and generalizable research ecosystem.

References

1. Key Statistics for ALL [Online], available at: https://www.cancer.org/cancer/types/acute-lymphocytic-leukemia/about/key-statistics.html (last accessed on 07/30/2023).
2. Cancer Stat Facts: ALL. [Online], available at: https://seer.cancer.gov/statfacts/html/alyl.html (last accessed on 07/30/2023).
3. SEER program [Online], available at: https://seer.cancer.gov/ (last accessed on 07/30/2023).
4. Nerella, Subhash, et al. "Transformers in Healthcare: A Survey." arXiv preprint arXiv:2307.00067 (2023).
5. Balabin, H. (2022). Multimodal Transformers for Biomedical Text and Knowledge Graph Data.
6. Thirunavukarasu, A.J., et al. Large language models in medicine. Nat Med (2023). https://doi.org/10.1038/s41591-023-02448-8.
7. Holmes, J., Liu, Z., et al. (2023). Evaluating large language models on a highly specialized topic, radiation oncology physics. arXiv preprint arXiv:2304.01938.
8. Wang, D. Q., Feng, et al. (2023). Accelerating the integration of ChatGPT and other large-scale AI models into biomedical research and healthcare. MedComm–Future Medicine, 2(2), e43.
9. Li, J., Dada, A., Kleesiek, J., & Egger, J. (2023). ChatGPT in Healthcare: A Taxonomy and Systematic Review. medRxiv, 2023-03.
10. Batzoglou, S. Large Language Models in Molecular Biology Deciphering the language of biology, from DNA to cells to human health. [Online], available at: https://towardsdatascience.com/large-language-models-in-molecular-biology-9eb6b65d8a30 (last accessed on 07/30/2023).
11. M. Aria, et al. ALL image dataset." Kaggle, (2021). DOI: 10.34740/KAGGLE/DSV/2175623.
12. Zhuofan Xia, et al. (2022) Vision Transformer with Deformable Attention, https://doi.org/10.48550/arXiv.2201.00520.
13. Swin transformer repo [Online], available at: https://github.com/microsoft/Swin-Transformer (last accessed on 07/30/2023).
14. M. Xu, K. Huang et al. "Multi-Task Learning with Context-Oriented Self-Attention for Breast Ultrasound Image Classification and Segmentation, doi: 10.1109/ISBI52829.2022.9761685.
15. K. Huang et al. "Shape-Adaptive Convolutional Operator for Breast Ultrasound Image Segmentation," 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 2021, pp. 1-6, doi: 10.1109/ICME51207.2021.9428287.
16. Jenny Li, et al. (2021). Evaluating Deep Learning Biases Based on Grey-Box Testing Results. IntelliSys 2020. Advances in Intelligent Systems and Computing, vol 1250. Springer, Cham, DOI=https://doi.org/10.1007/978-3-030-55180-3_48.
17. Tellez, N., Serra, J., Kumar. Y., et al. (2023). Gauging Biases in Various Deep Learning AI Models. IntelliSys 2022. Lecture Notes in Networks and Systems, vol 544. Springer, Cham. https://doi.org/10.1007/978-3-031-16075-2_11.
18. N. Tellez et al., "An Assure AI Bot (AAAI bot)," ISNCC, Shenzhen, China, 2022, pp. 1-5, doi: 10.1109/ISNCC55209.2022.9851759.
19. Kumar Y, et al. A Testing Framework for AI Linguistic Systems (testFAILS). Electronics. 2023; 12(14):3095. https://doi.org/10.3390/electronics12143095.
20. Pre-trained Faster R-CNN (Region-based Convolutional Neural Network) code, provided by ChatGPT-4 code interpreter [Online], available at: https://colab.research.google.com/drive/1DcdStV-xRzLzkCJ9d8eMVRRfTjSSfadH?usp=sharing (last accessed on 7/31/2023).
21. Y. Kumar, at al. (2023) AssureAIDoctor- A Bias-Free AI Bot. In proceeding of the 2023 International Symposium on Networks, Computers and Communications (ISNCC): Artificial Intelligence and Machine Learning. (ISNCC 2023)
22. Chat with ChatGPT-4-Turbo [Online], available at: https://chat.openai.com/share/4cfc5124-910b-49df-9e91-ad5e98fd91f6 (last accessed on 7/31/2023).
23. Selvaraju, Ramprasaath R. et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In Proceedings of the IEEE international conference on computer vision, pp. 618-626. 2017.