Research Statement

Nagur Shareef Shaik
https://scholar.google.com/citations?user=rkfPEjUAAAAJ

Research Interests: The complexities of human cognition remain one of the greatest mysteries of science. How can a network of neurons handle such complex tasks? From paying attention to minute details in noisy, distorted environments to perceiving and synthesizing information from diverse sources to understand intricate relationships. The ability to process and act on this information is truly awe-inspiring. My research interest centers around this mystery, aiming to unravel it through technology. I focus on artificial neural networks, emphasizing **attention mechanisms** and **multi-modal learning**, to enable machines to effectively process and synthesize images, text, and other modalities, driving intelligent decisions in complex, dynamic environments like medicine.

Research Experience: My research journey began during my undergraduate studies, where I focused on advancing representation learning with attention-based neural networks. I explored how models could attend to multi-level and multi-stage features across diverse imaging modalities, including retinal scans, brain MRIs, and chest CT scans, to diagnose diseases like diabetic retinopathy, brain tumors, and COVID-19. This work led to the development of composite mechanisms like hinge-attention networks and novel ensemble classifiers, which maximized margin in the latent feature space, significantly improving diagnostic accuracy and robustness [8], [10]–[14]. Building on this foundation, I expanded my research horizons as a Graduate Research Assistant at the TReNDS Center, where I systematically explored advanced AI techniques, tackling challenges that honed my expertise and opened new avenues for innovation.

1. **Attention Neural Networks:** One of the major challenges I faced early on was designing attention mechanisms tailored to high-dimensional neuroimaging and retinal data. The existing attention methods were insufficient in addressing the complexity of data from modalities like 3D structural MRI, which is critical for diagnosing schizophrenia. To tackle this, I developed the Spatial Sequence Attention Network, leveraging ConvLSTM to integrate spatial and volumetric patterns in the data. This approach led to a 6.52% improvement in performance [7]. Similarly, when working with retinal images, I found that conventional attention mechanisms struggled to capture both local and global contexts. To overcome this, I introduced Guided Context Gating, a method that uses global context features as guiding signals to refine local feature extraction through an attention gate. This innovation resulted in a 6.53% performance gain [5]. These experiences taught me the importance of designing modality- and task-specific attention mechanisms [2], and paved the way for their adaptation to multi-modal learning tasks, enhancing model robustness and effectiveness.

2. **Multi-modal Learning & Fusion:** As my research evolved, I began to focus more on multi-modal learning, which is particularly beneficial for complex diagnostic tasks like schizophrenia. One of the most exciting aspects of my work was integrating structural MRI (sMRI), functional network connectivity (FNC), and genetic SNPs to gain a more holistic view of the neuro-genetic interactions involved in schizophrenia. However, this presented several challenges, primarily due to the heterogeneous nature of these data sources and the complex correlations between them. To address this, I developed the Multi-Modal Imaging Genomics Transformer [3], which leveraged autoencoders to project multi-modal data into a unified latent space and used Transformers to implicitly fuse these representations. This approach resulted in a 2.12% improvement in diagnostic accuracy, showcasing the potential of carefully orchestrated multi-modal fusion techniques in uncovering neuro-genetic patterns. Despite these successes, there are still open challenges in balancing modality-specific contributions and aligning shared representations, motivating me to continue refining multi-modal learning approaches, especially when working with vision language models.

3. **Vision Language Models:** My work has also extended into medical image captioning, where I focused on integrating multi-modal retinal scans with clinical keywords to improve diagnostic reports, particularly for rare ophthalmological conditions [6]. This task presented new challenges, especially in aligning the diverse modalities—image and text—while maintaining their contextual integrity. To tackle these challenges, I developed the Multi-modal Medical Image Transformer [9], which used an implicit fusion strategy to integrate image and text embeddings effectively. Building on this foundation, I introduced Guided Context Self-Attention [1], which further refined the alignment and interaction between modalities. However, these approaches were not without limitations, particularly in their inability to dynamically adapt the modalities during fusion. To address this, I designed the Medical Vision Language Transformer [4], incorporating innovative Abstractor and Adaptor modules. These modules attend to each modality independently before dynamically adapting them during fusion, improving both computational efficiency and model performance

for diverse medical imaging modalities, including Chest X-rays. Currently, I am focused on learning disentangled modality-specific and shared representations using explicit conditioning in Multi-modal Variational Autoencoders (VAE).

Through this cross-disciplinary experience, I have contributed to cutting-edge research in instilling artificial cognition into machines for medical diagnostics. Specifically, I applied deep learning models to integrate and analyze multi-modal datasets—ranging from small to large-scale—for disease classification and report generation. This work has led to 14 journal publications & 3 conference presentations (ISBI, ICIP 2024), accumulating over 850 citations. Some of our ongoing projects are currently under review for ISBI, ICASSP, and AAAI 2025.

Research Scope: As an active reviewer for over 25 journals and having reviewed more than 35 articles, I have gained deep insights into the latest challenges and developments in AI for medicine. Despite significant breakthroughs in multi-modal learning, vision-language models, and attention mechanisms, their application in medicine still faces notable limitations. Large vision-language models like LLaVA-Med and MedCLIP have shown significant progress for medical applications. However, challenges remain in scaling these models for broader clinical use and often struggle with noisy, heterogeneous, and imbalanced data commonly encountered in real-world medical settings. Moreover, they also raise important cautionary questions about how we accurately measure—*evaluate* & *explain*—progress in the field. These limitations present several open research questions: How can we design adaptive, self-calibrating models that dynamically adjust to the heterogeneity and imbalance in real-world medical data? How can we build explainable AI systems that not only provide accurate predictions but also offer interpretable insights for clinicians, fostering trust and transparency? In the context of multi-modal learning, how can we develop fusion strategies that capture both shared and modality-specific information while ensuring computational efficiency and model scalability? What techniques can we employ to handle data scarcity and ensure robustness in environments where labeled medical data is limited or costly to obtain? How can we optimize AI models for deployment in resource-constrained settings, such as edge computing or mobile platforms, without compromising performance? Addressing these questions will form the foundation of a research agenda that aims to push the boundaries of AI in healthcare.

Research Publications

[1] T. K. Cherukuri, N. S. Shaik, J. D. Bodapati, and D. H. Ye, "Gcs-m3vlt: Guided context self-attention based multi-modal medical vision language transformer for retinal image captioning," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Under Review, IEEE, 2025.

[2] T. K. Cherukuri, N. S. Shaik, S. R. Yellu, and D. H. Ye, "Dynamic contextual attention network: Transforming spatial representations into adaptive insights for endoscopic polyp diagnosis," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, Under Review, IEEE, 2025.

[3] N. S. Shaik, T. K. Cherukuri, V. D. Calhoun, and D. H. Ye, "Multi-modal imaging genomics transformer: Attentive integration of imaging with genomic biomarkers for schizophrenia classification," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, Under Review, arXiv preprint arXiv:2407.19385, IEEE, 2025.

[4] N. S. Shaik, T. K. Cherukuri, and D. H. Ye, "Medvlt: Focus & fusion in vision language transformer yield expert precision in medical image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Under Review, Association for the Advancement of Artificial Intelligence, 2025.

[5] T. K. Cherukuri, N. S. Shaik, and D. H. Ye, "Guided context gating: Learning to leverage salient lesions in retinal fundus images," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2024.

[6] N. S. Shaik and T. K. Cherukuri, "Gated contextual transformer network for multi-modal retinal image clinical description generation," *Image and Vision Computing*, vol. 143, no. C, 2024.

[7] N. S. Shaik, T. K. Cherukuri, V. Calhoun, and D. H. Ye, "Spatial sequence attention network for schizophrenia classification from structural brain mr images," in *Proceedings of the 21st IEEE International Symposium on Biomedical Imaging*, IEEE, 2024.

[8] N. S. Shaik, T. K. Cherukuri, N. Veeranjaneulu, and J. D. Bodapati, "Medtransnet: Advanced gating transformer network for medical image classification," *Machine Vision and Applications*, vol. 35, no. 4, p. 73, 2024.

[9] N. S. Shaik, T. K. Cherukuri, and D. H. Ye, "M3t: Multi-modal medical transformer to bridge clinical context with visual insights for retinal image medical description generation," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2024.

[10] N. S. Shaik and T. K. Cherukuri, "Hinge attention network: A joint model for diabetic retinopathy severity grading," *Applied Intelligence*, vol. 52, no. 13, pp. 15 105–15 121, 2022.

[11] N. S. Shaik and T. K. Cherukuri, "Multi-level attention network: Application to brain tumor classification," *Signal, Image and Video Processing*, vol. 16, no. 3, pp. 817–824, 2022.

[12] N. S. Shaik and T. K. Cherukuri, "Transfer learning based novel ensemble classifier for covid-19 detection from chest ct-scans," *Computers in Biology and Medicine*, vol. 141, p. 105 127, 2022.

[13] J. D. Bodapati, N. S. Shaik, and V. Naralasetti, "Deep convolution feature aggregation: An application to diabetic retinopathy severity level prediction," *Signal, Image and Video Processing*, vol. 15, pp. 923–930, 2021.

[14] N. S. Shaik and T. K. Cherukuri, "Lesion-aware attention with neural support vector machine for retinopathy diagnosis," *Machine Vision and Applications*, vol. 32, no. 6, p. 126, 2021.