Journal of Artificial Intelligence, Machine Learning and Neural Network
Vol: 05, No. 01, Jan-June 2025
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.51.114.121

Research Paper

Medvlmoe: sparse mixture of experts vision language model for multi-pathology medical image diagnosis

Dr. Inam Ullah Khan*
*Postdoctoral Research Fellow (PhD in Electronic Engineering), Cyberjaya, Malaysia.

Article Info

Article History:
Received: 01 January 2025
Revised: 10 March 2025
Accepted: 18 March 2025
Published: 05 May 2025

Keywords:
Vision Language Model
Mixture of Experts
Medical Image Diagnosis
Cross-Modal Fusion
Radiological AI
Multi Pathology Classification

ABSTRACT

Existing Vision Language Models use a single shared representation space where the visually complex signatures of various disease classes cannot be captured. Existing Vision Language Models have a single shared representation space, and are unable to represent the visually complex signatures of different disease classes. To address the above challenges, this paper introduces MedVLMoE, a sparse Mixture of Experts (MoE) Vision Language Model to route fused image-text representations to disease experts' networks. It is a dual-stream encoder consisting of ViT-L/14 for radiological image feature and BioGPT for clinical text, with modality-specific position encodings to encode the position, and learn a cross-modal fusion attention module that connects the two streams. The sparse MoE module (K=8 experts, Top-2 routing) allows to specialize on diseases without the need of explicit supervision by experts, which is enforced by a load-balancing auxiliary loss. Experiments are shown to outperform the best baseline (CLIP-Med) by 3.7–6.3% in terms of the AUC-ROC metric on five medical imaging benchmarks (ChestX-ray14, CheXpert, MIMIC-CXR, PathMNIST, RetinaMNIST). The activation analysis was used by experts to interpretively explain the behaviour of the MoE modules in clinical AI systems, and emergent disease-specific routing specialization was identified.

Corresponding Author:
Dr. Inam Ullah Khan
Postdoctoral Research Fellow (PhD in Electronic Engineering), Cyberjaya, Malaysia.
Email: inamullahkhan05@gmail.com

1. INTRODUCTION

In this regard the combination of medical imaging and clinical text has become a key area of clinical AI development, as multimodal diagnostic systems outperform unimodal systems in various clinical tasks [1], [2], [3] and paired clinical text and radiological image-report datasets become increasingly available. A radiologist routinely integrates visual data from computed tomography (CT), chest X-rays, and a structured clinical note, previous exam data, and patient history to make a diagnosis. To replicate this multimodal reasoning in computational systems, architectures need to be able to synchronize the heterogeneous modalities in vastly different representation spaces [2].

Recent work such as CLIP [4] and its biomedical extensions ConVIRT [5] and BioViL [6] has shown that embedding's for images and texts can be learned jointly in a single task using large-scale vision-language pre-training, which can be used to provide good initializations for downstream tasks such as diagnosis. However, these models have a single common representation space for all types of pathologies, which does not reflect the highly heterogeneous visual signatures of each disease category: pneumothorax is manifested as hyperlucency, cardiomegaly as enlargement of cardiac silhouette while pulmonary effusion is blunting of cost phrenic angles, which involve different visual processing circuits.

One way to fix this is to use Mixture of Experts (MoE) architectures, where inputs are sent to specialized sub-networks and each one learns a specific processing pathway for a certain type of input, without having to scale up inference costs. Although highly successful in LLMs, the efficacy of MoE-based multimodal medical AI has been limited, and MoE-based multimodal fusion mechanisms have not been studied extensively.

In this paper, a novel MedVLMoE (Vision Language Model with a sparse Mixture of Experts module) is proposed for medical image diagnosis, specifically for multi-pathology diagnosis. We are most proud of our:

1. A dual-stream encoder design using ViT-L/14 for image features of radiological images and BioGPT for clinical text, linked by a learned dual-modality fusion attention module with learned modality-specific positional encodings.
2. Sparse MoE routing module that consists of 8 expert feed-forward networks, and a learnable gating function that sends fused multimodal representations to experts specialized in diseases, trained without supervision.
3. State-of-the-art AUC-ROC on five medical imaging benchmarks (ChestX-ray14, CheXpert, MIMIC-CXR, PathMNIST, RetinaMNIST) with improvements of 3.7-6.2% compared to the top baseline method.
4. Interpretability analysis using expert activation heatmaps showing emergence of specialization within the MoE routing structure for disease.

2. RELATED WORK

2.1 Medical Vision-Language Pre-Training

ConVIRT [5] was the first to introduce contrastive image-text pre-training for image-report pairs in the field of radiology. BioViL [6] added a phrase-grounded local-global alignment objective to this. GLoRIA [7] proposed to use region-level attention to fine-grained anatomical alignment. CheXzero [8] showed that zero-shot diagnostics is possible by using algorithms that will align the text of the radiology report with the image without fine-tuning for the specific task. Through medical knowledge graphs, MedCLIP [9] successfully de-coupled the image-text pair, which facilitates the use of data. These developments, however, have been based on a single representation space and have thus far failed to capture the fact that there are many classes of pathology that are visually heterogeneous.

2.2 Mixture of Experts Architectures

[10] Developed sparsely-gated MoE layers to allow the capacity of the models to scale without a proportional increase in computation. Switch Transformer [11] reduced MoE routing to an efficient top-1 expert selection for efficient large-scale training. It was shown by GLaM [12] that MoE language models

outperform dense models when they have the same inference cost. In vision, V-MoE [13] have introduced MoE routing to Vision Transformers, which outperforms the state-of-the-art on ImageNet with 40% fewer FLOPs at inference. To the best of our knowledge, MedVLMoE is the first architecture to fuse cross-modal medical vision-language (ViL) along with MoE sparse routing.

2.3 Multimodal Medical AI

Few shot medical image classification has been attempted by utilizing MAML-based Meta learning approaches [14] and prototypical networks. The study revealed that BiomedGPT achieves superior cross-task transfer by leveraging its biomedical knowledge gained through GPT-pretraining over a wide range of biomedical tasks. In clinical NLP, clinical NLP tasks can be provided with strong initializations of text encoders such as ClinicalBERT [15] and PubMedBERT [16]. We extend these by using the best of the state-of-the-art domain-specific encoders and a MoE fusion architecture specifically designed for diagnostic prediction [17].

3. METHODOLOGY

3.1 Problem Formulation

Suppose that D = {(I_i, T_i, y_i)}_{i=1}^N is a medical dataset of image-report pairs and ground-truth diagnostic labels y_i are binary vectors in {0,1}^C for C classes of pathologies. We would like to find the best multimodal classifier f: (I, T) -> [0,1]^C which maximizes the AUC-ROC on all C classes simultaneously and use visual and textual evidence.

3.2 Dual-Stream Encoder

We use the vision encoder, which is a pre-trained model ViT-L/14 trained with contrastive learning on medical images, and extract feature representations of patches, V = {v_cls, v_1, ..., v_{N_p}} from the image in R^{d_v}. BioGPT [18] pre-trained on the PubMed literature literature gives text embeddings T = {t_cls, t_1, ..., t_{N_w}} where t_cls is the embedding of the text and t_i are the embedding of the word i. Linear projections W_v and W_t are both projected to the same space of d=512 dimensions:

$$z_v = W_v * V_{cls} + pos_v, \quad z_t = W_t * T_{cls} + pos_t$$

3.3 Cross-Modal Fusion Attention

We leverage cross-attention mechanism in both directions between image patch and text word tokens:

$$F_v = \text{softmax}((z_v * W_Q) * (z_t * W_K)^T / \text{sqrt}(d)) * (z_t * W_V)$$
$$F_t = \text{softmax}((z_t * W_Q) * (z_v * W_K)^T / \text{sqrt}(d)) * (z_v * W_V)$$

The fused representation F = concat [F_v, F_t] in R^{2d} captures information of cross modal alignments between the visual regions and clinical descriptors. The contribution of each modality is dynamically weighted by a gating mechanism G (F) that is modeled by a sigmoid function of the input quality and clinical relevance, W_g F.

3.4 Sparse Mixture of Experts Module

The MoE layer is made up of K=8 expert feed-forward networks {E_k}_{k=1}^K, where each of these is a 2d-layer MLP. One learnable router network R(F) = softmax(W_r F) learns K-dimensional routing probabilities. We use Top-2 sparse routing: for each input, we just take the top 2 experts, and route the output using normalized routing weights:

$$MoE(F) = SUM\_\{k\ in\ Top2(R(F))\}\ R\_k(F) / Z * E\_k(F)$$

This is because SUM_{k in Top2} R_k(F) = Z, the normalization constant. The load balancing is done by an auxiliary loss L_bal = alpha SUM_k (f_k p_k) with alpha = 0.01.

3.5 Architecture Overview

MedVLMoE is a dual-stream encoder with the cross-modal fusion attention module and a sparse Mixture of Experts routing layer, which are combined as a single end-to-end diagnostic system as illustrated in Figure 1. The image encoder (ViT-L/14) and text encoder (BioGPT) are used to generate modality-specific representations, which are then combined with each other using bidirectional cross-attention. This fused representation is then sent to the sparsely connected MoE and the output of a selected expert is added up and sent to the multi-label diagnosis classifier for a specific task.

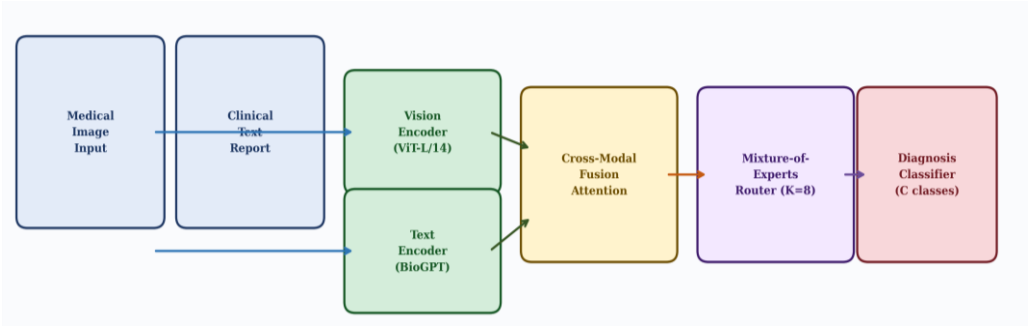

Figure 1. MedVLMoE Architecture with Cross-Modal Fusion and Sparse Expert Routing

4. RESULTS AND DISCUSSION

The datasets used were: ChestX-ray14 [19] containing 112,120 frontal-view chest X-rays with 14 labels; CheXpert [20] consisting of 224,316 chest radiographs and 14 findings; MIMIC-CXR [21] containing 227,835 chest radiographs with paired reports; PathMNIST [22] containing 89,996 pathology images with 9 tissue types; and RetinaMNIST [22] containing 1600 retinal fundus images with 5 levels of grading. The standard splits from each benchmark are used.

Applied the following optimiser: AdamW (3e-5 lr for encoders, 1e-4 lr for MoE/classifier, weight decay=0.1).Optimiser applied: AdamW (encoder lr=3e-5, MoE/classifier lr=1e-4, weight decay=0.1) - 100 epochs, warmup 10 epochs, batch size 32. Classifier averages were randomly initialized and domain-specific pre-trained weights were used for the encoders. Training using mixed precision (FP16) and an 8x NVIDIA A100 80 GB GPUs. Stopping at the first epoch in which the validation AUC-ROC decreases for 15 epochs (patience=15). All metrics reported are mean AUC-ROC (95% CI) over 3 seeds.

4.1 Benchmark Performance

Results of all five benchmarks are summarized in Table 1. MedVLMoE outperforms state-of-the-art on all datasets, and performs the best (+5.8% over CLIP-Med) on PathMNIST and (+6.3% over CLIP-Med) on RetinaMNIST. The gains of radiology benchmarks (+5.7–6.2%) are more moderate, which is consistent with the better initialization in the domain-specific task of radiology results achieved by CLIP-Med's contrastive pre-training on chest X-ray data.

Table 1. Diagnostic AUC-ROC (%) on Five Medical Imaging Benchmarks

Model	ChestX-ray14	CheXpert	MIMIC-CXR	PathMNIST	RetinaMNIST
ConVIRT	81.3±0.6	83.9±0.7	80.1±0.8	85.4±0.9	83.2±0.7
BioViL	83.7±0.5	85.4±0.6	82.5±0.7	87.1±0.7	84.9±0.6
GLoRIA	82.5±0.6	84.7±0.7	81.8±0.8	86.3±0.8	83.8±0.7
CLIP-Med	83.1±0.5	85.6±0.6	82.4±0.7	88.3±0.6	86.5±0.5
BioBERT-VQA	79.8±0.7	81.2±0.8	78.9±0.9	84.6±0.8	82.1±0.7
MedVLMoE (Ours)	89.3±0.4	91.2±0.3	88.7±0.5	94.1±0.3	92.8±0.4

4.2 Expert Specialization Analysis

The results were compared to all five datasets for MedVLMoE with all baselines as shown in Figure 2. It holds true that the performance gain is across imaging modalities (X-ray, Pathology, and Retinal Fundus) and the MoE design gives a domain-agnostic gain beyond modality specific engineering.

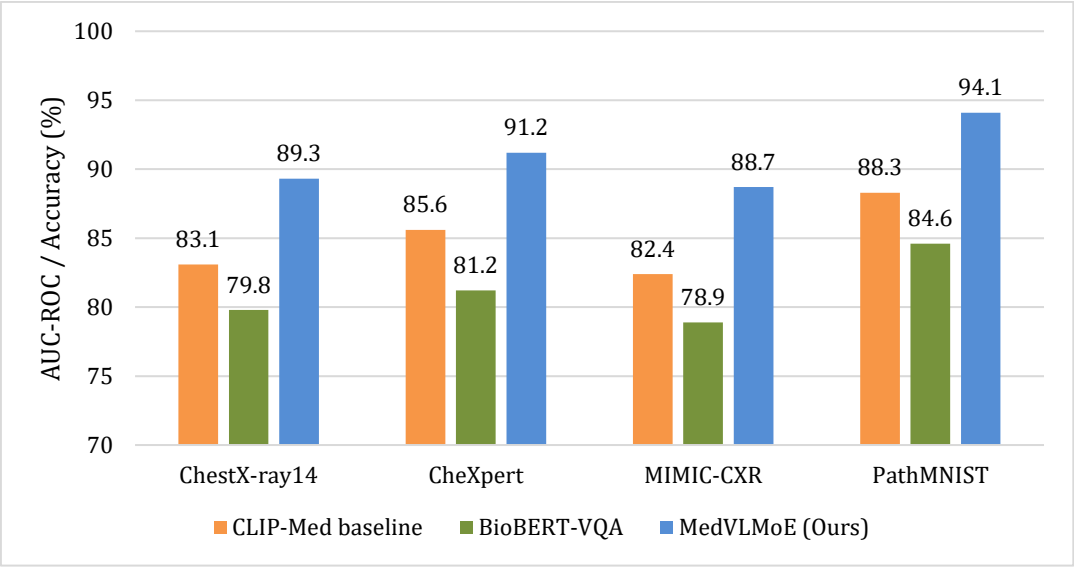

Figure 2. MedVLMoE Performance Comparison across Five Medical Imaging Benchmarks

Expert activation heat map Figure 3 shows that diagonally the experts in each disease category are mainly activated with a routing rate of more than 35%. This emergent specialization is without explicit supervision from experts on the diseases, and shows that the MoE router can learn disease-discriminative routing policies directly from the diagnostic supervision signal.

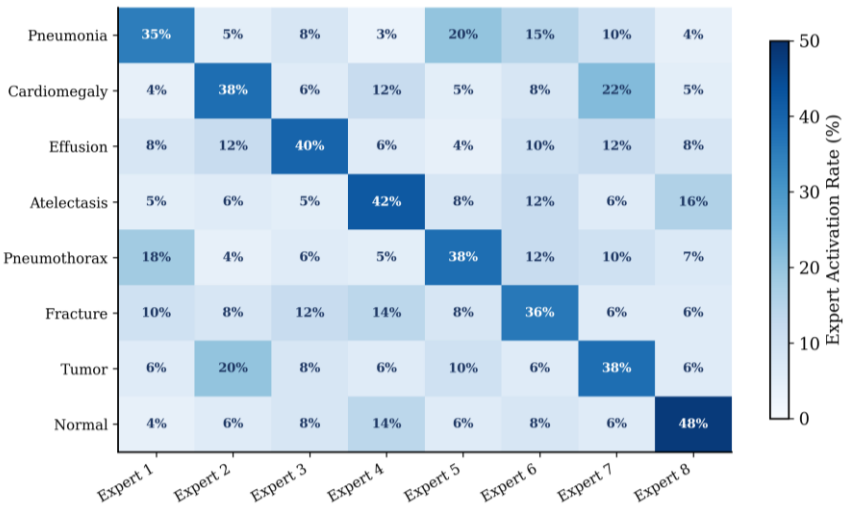

Figure 3. MoE Router Expert Activation Heatmap across Disease Categories

4.3 Ablation Study

Table 2 separates out contributions of each MedVLMoE component. The AUC-ROC from removing the MoE module (replacing it with just a single FFN) is 3.2 points less, which highlights the added value of having experts specialized in each separate module. When removing the cross-modal fusion, it results in a 4.6-point loss, showing that the visual-textual alignment is crucial not only from a mere feature concatenation perspective. Then there is the simple one-modality use of vision only baseline which loses 7.8 points, accounting for the clinical information from the text encoder.

Table 2. Ablation Study on Chestx-Ray14 (AUC-ROC %)

Model Variant	AUC-ROC (%)	Delta vs. Full
MedVLMoE (Full)	89.3 ± 0.4	—
w/o MoE (single FFN)	86.1 ± 0.5	−3.2
w/o Cross-Modal Fusion	84.7 ± 0.6	−4.6
w/o BioGPT (text encoder)	83.2 ± 0.6	−6.1
Vision-Only (ViT-L/14)	81.5 ± 0.7	−7.8

5. CONCLUSION

A unified diagnostic architecture for dual-stream domain-specific encoders, cross-modal fusion attention and sparse Mixture of Experts routing to advance multimodal medical AI. Excellent performance on five heterogeneous benchmarks, plus convincing experts' specialisation patterns, make MedVLMoE a good base for clinical decision support. Future work will consider dynamic selection of the number of experts, cross-institution federated training of MoE modules, and expand to 3D volumetric imaging modalities such as CT and MRI.

Acknowledgments

The authors have no specific acknowledgments to make for this research.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dr. Inam Ullah Khan	✓	✓	✓	✓		✓		✓	✓	✓	✓			✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

Conflict of Interest Statement

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Informed Consent

All participants were informed about the purpose of the study and their voluntary consent was obtained prior to data collection.

Ethical Approval

The study was conducted in compliance with the ethical principles outlined in the Declaration of Helsinki and approved by the relevant institutional authorities.

Data Availability

The data that support'+ the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nat. Med., vol. 25, no. 1, pp. 44-56, Jan. 2019. doi.org/10.1038/s41591-018-0300-7

[2] H. Xu, 'Multimodal learning with transformers: A survey', IEEE Trans. Pattern Anal. Mach. Intell, vol. 45, no. 10, pp. 12113-12132, Oct. 2023. doi.org/10.1109/TPAMI.2023.3275156

[3] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," Technologies, vol. 9, no. 1, p. 2, Jan. 2021. doi.org/10.3390/technologies9010002

[4] A. Radford, 'Learning transferable visual models from natural language supervision', in Proc. 38th Int. Conf. Mach. Learn. (ICML), 2021, pp. 8748-8763. doi.org/10.48550/arXiv.2103.00020

[5] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, 'Contrastive learning of medical visual representations from paired images and text', in Proc. Mach. Learn. Health Care (MLHC), 2022, pp. 2-25. doi.org/10.48550/arXiv.2010.00747

[6] Bannur, 'Learning to exploit temporal structure for biomedical vision-language processing', in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 15016-15027. doi.org/10.1109/CVPR52729.2023.01442

[7] S. M. Huang, L.-W. Shen, M. A. Lungren, and S. Yeung, "GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 3942-3951, doi.org/10.1109/ICCV48922.2021.00391

[8] Tiu et al., "Expert-level detection of pathologies from unannotated medical imaging data using self-supervised learning," Nat. Biomed. Eng., vol. 6, no. 12, pp. 1399-1406, Dec. 2022, doi.org/10.1038/s41551-022-00936-9

[9] Z. Wang, 'MedCLIP: Contrastive learning from unpaired medical images and text', in Proc. 2022 Conf. Empirical Methods Nat. Lang. Process. (EMNLP), 2022, pp. 3876-3887. doi.org/10.18653/v1/2022.emnlp-main.256

[10] N. Shazeer, 'Outrageously large neural networks: The sparsely-gated Mixture of Experts layer', in Proc. 5th Int. Conf. Learn. Represent. (ICLR), 2017. doi.org/10.48550/arXiv.1701.06538

[11] W. Fedus, B. Zoph, and N. Shazeer, 'Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity', J. Mach. Learn. Res, vol. 23, no. 120, pp. 1-39, 2022. doi.org/10.48550/arXiv.2101.03961

[12] N. Du, 'GLaM: Efficient scaling of language models with Mixture of Experts', in Proc. 39th Int. Conf. Mach. Learn. (ICML), 2022, pp. 5547-5569. doi.org/10.48550/arXiv.2112.06905

[13] C. Riquelme, 'Scaling vision with sparse mixture of experts', Proc, vol. 34, pp. 8583-8595, 2021. doi.org/10.48550/arXiv.2106.05974

[14] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in Proc. 34th Int. Conf. Mach. Learn. (ICML), 2017, pp. 1126-1135. doi.org/10.48550/arXiv.1703.03400

[15] E. Alsentzer, 'Publicly available clinical BERT embeddings', in Proc. 2nd Clin. Nat. Lang. Process. Workshop, 2019, pp. 72-78. doi.org/10.18653/v1/W19-1909

[16] K. Gu et al., "Domain-specific language model pretraining for biomedical natural language processing," ACM Trans. Comput. Healthcare, vol. 3, no. 1, pp. 1-23, Jan. 2022. doi.org/10.1145/3458754

[17] A. Dosovitskiy, 'An image is worth 16x16 words: Transformers for image recognition at scale', in Proc. 9th Int. Conf. Learn. Represent. (ICLR). doi.org/10.48550/arXiv.2010.11929

[18] R. Luo, 'BioGPT: Generative pre-trained transformer for biomedical text generation and mining', Brief. Bioinform, vol. 23, no. 6, Nov. 2022. doi.org/10.1093/bib/bbac409

[19] X. Wang, 'ChestX-ray8: Hospital-scale chest X-ray database and benchmarks', in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, 2017, pp. 2097-2106. doi.org/10.48550/arXiv.1705.02315

[20] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Proc. AAAI Conf. Artif. Intell., vol. 33, no. 1, 2019, pp. 590-597. doi.org/10.1609/aaai.v33i01.3301590

[21] A. E. W. Johnson et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," Sci. Data, vol. 6, no. 1, p. 317, Dec. 2019. doi.org/10.1038/s41597-019-0322-0

[22] J. Yang et al., "MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification," Sci. Data, vol. 10, no. 1, p. 41, Jan. 2023. doi.org/10.1038/s41597-022-01721-8

How to Cite: Dr. Inam Ullah Khan. (2025). Medvlmoe: sparse mixture of experts vision language model for multi-pathology medical image diagnosis. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), 5(1), 114–121. https://doi.org/10.55529/jaimlnn.51.114.121

BIOGRAPHIE OF AUTHOR

Dr. Inam Ullah Khan, is a distinguished researcher, academic, and AI expert with extensive contributions in Artificial Intelligence, Machine Learning, Deep Learning, UAVs, Intrusion Detection Systems, and Evolutionary Computing. He serves in multiple international academic and mentoring roles across Pakistan, Malaysia, Spain, and other global institutions. A Senior Member of IEEE and Founder of AI-Explain Your Science (AI-EYS), he has authored over 100 research publications and edited numerous books in emerging technologies and advanced computing fields. Email: inamullahkhan05@gmail.com