

---

**REFERENCE**

1. Bazi, Yakoub, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. 2023. "Vision–Language Model for Visual Question Answering in Medical Imagery" *Bioengineering* 10, no. 3: 380.
2. Claudio Filipi Goncalves dos Sants, Felype de Castro Bastos, Ana Claudia Akemi Matsuki de Faria et al. "Visual Question Answering: A Survey on Techniques and Common Trends in Recent Literature", 05 June 2023, PREPRINT (Version 1) available at Research Square [<https://doi.org/10.21203/rs.3.rs-3015858/v1>]
3. Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Yang Feng, Jin Hao, Junhui Lv, Zuozhu Liu. "Parameter-Efficient Transfer Learning for Medical Visual Question Answering", *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023.
4. Zhihong Lin, Donghao Zhang, Qingyi Tac, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. "Medical visual question answering: A survey", 2022.
5. Lu S, Liu M, Yin L, Yin Z, Liu X, Zheng W. "The multi-modal fusion in visual question answering: a review of attention mechanisms". *PeerJ Comput Sci.* 2023 May 30;9:e1400. doi: 10.7717/peerj-cs.1400. PMID: 37346665; PMCID: PMC10280591.
6. V. Kodali and D. Berleant, "Recent, Rapid Advancement in Visual Question Answering: a Review," 2022 IEEE International Conference on Electro Information Technology (EIT), 2022, pp. 139-146.
7. Sruthy Manmadhan and Binsu C. Kavoov, "Visual question answering:a state-of-the-art review," *Artif. Int. Rev.*, vol. 53, pp. 5705-5745, 2020, <https://doi.org/10.1007/s10462-020-09832-7>.
8. Himanshu Sharma and Anand Singh Jalal. "A survey of methods, datasets and evaluation metrics for visual question answering. *Image and Vision Computing*", 116:104327, 2021.
9. Charulata Patil and Manasi Patwardhan. "Visual question generation: The state of the art". *ACM Comput. Surv.*, 53(3), may 2020.

10. Yeyun Zou and Qiyu Xie. "A survey on VQA: Datasets and approaches". In 2020 2nd International Conference on Information Technology and Computer Application (ITCA). IEEE, dec 2020.
11. Fuji Ren and Yangyang Zhou, "CGMVQA a new classification and generative model for medical visual question answering", IEEE Access, vol. 8, pp. 50626-50636, 2020.
12. Lubna A, Saidalavi Kalady and Lijiya A, "MoBVQA a modality based medical image visual question answering system", TENCON 2019 - 2019 IEEE Region 10 Conference IEEE, 17-20 October 2019, Kochi, India, 2019.
13. Fazal Muhammad, Ziaul Haq Abbas, Ghulam Abbas and Lei Jiao, "Decoupled downlink-uplink coverage analysis with interference management for enriched heterogeneous cellular networks", IEEE Access, vol. 4, pp. 6250-6260, 2016.
14. Dhruv Sharma, Sanjay Purushotham and Chandan K Reddy, "MedFuseNet an attention based multimodal deep learning model for visual question answering in the medical domain", Scientific Reports, vol. 11, no. 1, pp. 1-18, 2021.
15. Shengyan Liu, Xuejie Zhang, Xiaobing Zhou and Jian Yang, "BPI-MVQA a bi-branch model for medical visual question answering", BMC Medical Imaging, vol. 22, no. 1, pp. 1-19, 2022.
16. Yakoub Bazi 1, Mohamad Mahmoud Al Rahhal 2, Laila Bashmal 1 and Mansour Zuair 1 "Vision–Language Model for Visual Question Answering in Medical Imagery", Bioengineering 2023.
17. Li, L.; Lei, J.; Gan, Z.; Liu, J. Adversarial "VQA: A New Benchmark for Evaluating the Robustness of VQA Models". In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 2022–2031.
18. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding", arXiv 2019, arXiv:1810.04805.
19. He K, Zhang X, Ren S, Sun J. "Deep residual learning for image recognition". In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016;770–778.

- 
20. Srinivasan K, Garg L, Datta D, Alaboudi AA, Jhanjhi NZ, Agarwal R, Thomas AG. "Performance comparison of deep cnn models for detecting driver's distraction". *CMC-Comput Mater Continua*. 2021;68(3):4109–24.
  21. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. "Learning phrase representations using rnn encoder-decoder for statistical machine translation", arXiv preprint arXiv: 1406. 1078, 2014.
  22. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. "Biobert: a pre-trained biomedical language representation model for biomedical text mining". *Bioinformatics*. 2020;36(4):1234–40.
  23. Devlin J, Chang M-W, Lee K, Toutanova K. "Bert: pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv: 1810. 04805, 2018.
  24. Peng Y, Liu F, Rosen MP. "Umass at imageclef medical visual question answering (med-vqa) 2018 task". In *CLEF (Working Notes)*, 2018.
  25. Zhejiang University at CLEF Image Retrieval and Classification Task 2019 'Visual Question Answering in the Medical Domain. 2019'.
  26. Kornuta T, Rajan D, Shivade C, Asseman A, Ozcan AS. "Leveraging medical visual question answers with supporting facts", arXiv preprint arXiv: 1905. 12008, 2019.
  27. Liao Z, Wu Q, Shen C, Van Den Hengel A, Verjans J. "Aiml at Healthcare Image QA 2020: Knowledge inference via a skeleton-based sentence mapping approach for medical domain visual question answering", 2020.
  28. Al-Sadi A, Hana'Al-Theiabat, Al-Ayyoub M. "The inception team at Healthcare Image QA 2020: Pretrained vgg with data augmentation for medical vqa and vqg". In *CLEF (Working Notes)*, 2020.
  29. Zhan L-M, Liu B, Fan L, Chen J, Wu X-M. "Medical visual question answering via conditional reasoning". In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020;2345–2354.
  30. Abacha, A.B., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: "Healthcare Image QA: Overview of the medical visual question answering task at CLEF Image Retrieval and Classification Task 2019". In: *CLEF (Working Notes)* (2019)
-

- 
- 
31. Y. I. Jinesh Melvin, S. Gawade and H. Palivela, "Visual Question Answering using Data Mining Techniques for Skeletal Scintigraphy in medical domain - VQADMSS," *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Coimbatore, India, 2021, pp. 859-863, doi: 10.1109/ICAIS50930.2021.9395936.
  32. Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. "Towards vqa models that can read". In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8309–8318, 2019.
  33. Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluís Gomez, Marçal Rusinol, Ernest Valveny, C.V. Jawahar, and Dimosthenis Karatzas. "Scene text visual question answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)", October 2019.
  34. Chao Yang, Su Feng, Dongsheng Li, Huawei Shen, Guoqing Wang, and Bin Jiang. "Learning content and context with language bias for visual question answering". In 2021 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6, July 2021.
  35. Dirk Vath, Pascal Tilli, and Ngoc Thang Vu. "Beyond accuracy: A consolidated tool for visual question answering benchmarking". In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 114–123, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
  36. A. Farinhas, A. T. Martins, and P. Q. Aguiar. "Multimodal continuous visual attention mechanisms". In 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pages 1047–1056, Los Alamitos, CA, USA, Oct 2021. IEEE Computer Society.
  37. Corentin Kervadec, Grigory Antipov, "Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should vqa expect them to"? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2776–2785, 2021.

38. D. Teney, E. Abbasnejad, and A. van den Hengel. “Unshuffling data for improved generalization in visual question answering”. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 1397–1407, Los Alamitos, CA, USA, Oct 2021. IEEE Computer Society.
39. Chenyu Gao, Qi Zhu, Peng Wang, Hui Li, Yuliang Liu, Anton Van den Hengel, and Qi Wu. “Structured multimodal attentions for textvqa. IEEE Transactions on Pattern Analysis and Machine Intelligence”, 2021.
40. Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z Pan, Zonggang Yuan, and Huajun Chen. “Zero-shot visual question answering using knowledge graph”. In International Semantic Web Conference, pages 146–162. Springer, 2021.
41. Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. “Medical visual question answering via conditional reasoning”. In Proceedings of the 28th ACM International Conference on Multimedia, MM ’20, page 2345–2354, New York, NY, USA, 2020. Association for Computing Machinery.
42. Vatsal Goel, Mohit Chandak, Ashish Anand, and Prithwijit Guha. “Iq-vqa: Intelligent visual question answering”. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani, editors, Pattern Recognition. ICPR International Workshops and Challenges, pages 357–370, Cham, 2021. Springer International Publishing.
43. S. Whitehead, H. Wu, H. Ji, R. Feris, and K. Saenko. “Separating skills and concepts for novel visual question answering”. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5628–5637, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society.
44. Zhiqian Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu. “Debiased visual question answering from feature and sample perspectives”. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, 2021.
45. Rajat Koner, Hang Li, Marcel Hildebrandt, Deepan Das, Volker Tresp, and Stephan Günnemann. Graphhopper: “Multi-hop scene graph reasoning for visual question answering”. In Andreas Hotho, Eva Blomqvist, Stefan Dietze, Achille Fokoue, Ying Ding, Payam Barnaghi, Armin Haller, Mauro Dragoni,

- 
- 
- and Harith Alani, editors, *The Semantic Web – ISWC 2021*, pages 111–127, Cham, 2021. Springer International Publishing.
46. Claudio Greco, Barbara Plank, Raquel Fernández, and Raffaella Bernardi. “Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering”. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3601–3605, Florence, Italy, July 2019. Association for Computational Linguistics.
  47. Junjie Wang, Yatai Ji, Jiaqi Sun, Yujiu Yang, and Tetsuya Sakai. “Mirtt: Learning multimodal interaction representations from trilinear transformers for visual question answering”. pages 2280–2292, 01 2021.
  48. Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. “Contrastive Pre-training and Representation Distillation for Medical Visual Question Answering Based on Radiology Images”, pages 210–220. 09 2021.
  49. Kushal Kafle, Robik Shrestha, Scott Cohen, Brian Price, and Christopher Kanan. “Answering questions about data visualizations using efficient bimodal fusion”. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pages 1498–1507, 2020.
  50. Binh X Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D Tran, and Anh Nguyen. “Coarse-to-fine reasoning for visual question answering”. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4558–4566, 2022.
  51. Haifan Gong, Ricong Huang, Guanqi Chen, and Guanbin Li. “Sysu-hcp at Healthcare Image QA 2021: A data-centric model with efficient training methodology for medical visual question answering”. *Proceedings* <http://ceur-ws.org> ISSN, 1613:0073, 2021.
  52. Anwen Hu, Shizhe Chen, and Qin Jin. “Question-controlled text-aware image captioning”. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3097–3105, 2021.
  53. Yibing Liu, Yangyang Guo, Jianhua Yin, Xuemeng Song, Weifeng Liu, Liqiang Nie, and Min Zhang. “Answer questions with right image regions: A visual attention regularization approach”. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(4):1–18, November 2022.
- 
-

54. Leonard Salewski, A. Sophia Koepke, Hendrik P. A. Lensch, and Zeynep Akata. CLEVR-x: “A visual reasoning dataset for natural language explanations”. In *xxAI- Beyond Explainable AI*, pages 69–88. Springer International Publishing, 2022.
55. Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. “VLMo: Unified vision-language pre-training with mixture-of-modality-experts”. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
56. Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. “Greedy gradient ensemble for robust visual question answering”. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2021.
57. Zujie Liang, Haifeng Hu, and Jiaying Zhu. “LPF: A language-prior feedback objective function for de-biased visual question answering”. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, July 2021.
58. Chen Qu, Hamed Zamani, Liu Yang, W. Bruce Croft, and Erik Learned-Miller. “Passage retrieval for outside-knowledge visual question answering”. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, July 2021.
59. Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R. Manmatha. “Latr: Layout-aware transformer for scene-text vqa”. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16527–16537, June 2022.
60. Emanuele Vivoli, Ali Furkan Biten, Andres Mafla, Dimosthenis Karatzas, and Lluís Gomez. “Must-vqa: Multilingual scene-text vqa”. In *Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, page 345–358, Berlin, Heidelberg, 2023. Springer-Verlag.
61. C. Kervadec, T. Jaunet, G. Antipov, M. Baccouche, R. Vuillemot, and C. Wolf. “How transferable are reasoning patterns in vqa?”. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4205–4214, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society.

62. Aisha Urooj Khan, Hilde Kuehne, Chuang Gan, Niels da Vitoria Lobo, and Mubarak Shah. “Weakly supervised grounding for vqa in vision-language transformers”. 2022.
63. Herring W, Learning radiology: “Recognizing the basics. Elsevier Health Sciences”, 2015.
64. Y. I. Jinesh Melvin, Sushopti Gawade, Hemant Palivela, "Feature Extraction from Radiology Images for Visual Question Answering System Using CNN and BiLSTM Model", Recent Innovations in Computing, vol.832, pp.317, 2022.
65. Novelline RA and Squire LF, “Squire’s fundamentals of radiology”. La Editorial, UPR, 2004.
66. Jones J, Normal abdominal x-ray. Case study, Radiopaedia.org (Accessed on 01 Feb 2024) <https://doi.org/10.53347/rID-34067>.
67. Hasan, S.A., Ling, Y., Farri, O., Liu, J., Müller, H., Lungren, M.P., 2018. “Overview of ImageCLEF 2018 medical domain visual question answering task”, in: CLEF (Working Notes).
68. Lau, J.J., Gayen, S., Abacha, A.B., Demner-Fushman, D., 2018. “A dataset of clinically generated visual questions and answers about radiology images”. Scientific Data 5, 1–10.
69. Ben Abacha, A., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H., 2019. Healthcare Image QA: Overview of the medical visual question answering task at CLEF Image Retrieval and Classification Task 2019, in: CLEF2019 Working Notes, CEUR-WS.org, Lugano, Switzerland.
70. Kovaleva, O., Shivade, C., Kashyap, S., Kanjaria, K., Wu, J., Ballah, D., Coy, A., Karargyris, A., Guo, Y., Beymer, D.B., et al., 2020. “Towards visual dialog for radiology”, in: Proceedings of the 19th SIG-BioMed Workshop on Biomedical Language Processing, pp. 60–69.
71. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P., 2020. “PathVQA: 30000+ questions for medical visual question answering”. arXiv preprint arXiv:2003.10286 .
72. Ben Abacha, A., Datla, V.V., Hasan, S.A., Demner-Fushman, D., Müller, H., 2020. “Overview of the Healthcare Image QA task at ImageCLEF 2020:

- 
- Visual question answering and generation in the medical domain”, in: CLEF 2020 Working Notes, CEUR-WS.org, Thessaloniki, Greece.
73. Ben Abacha, A., Sarrouiti, M., Demner-Fushman, D., Hasan, S.A., Müller, H., 2021. “Overview of the Healthcare Image QA task at ImageCLEF 2021: Visual question answering and generation in the medical domain”, in: CLEF 2021 Working Notes, CEUR-WS.org, Bucharest, Romania.
  74. Liu, S., Ou, X., Che, J., Zhou, X., Ding, H., 2019. “An Xception GRU model for visual question answering in the medical domain”, in: CLEF (Working Notes).
  75. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D., 2017. “Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering”, in: Conference on Computer Vision and Pattern Recognition (CVPR).
  76. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. “Microsoft COCO: Common objects in context”, in: European conference on computer vision, Springer. pp. 740–755.
  77. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S., 2019. “MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs”. arXiv preprint arXiv:1901.07042 .
  78. Liu, B., Zhan, L.M., Wu, X.M., 2021a. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), “Medical Image Computing and Computer Assisted Intervention – MICCAI 2021”, Springer International Publishing, Cham. pp. 210–220.
  79. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M., Golia-Pernicka, J., Heckers, S.H., Jarnagin, W.R., McHugo, M.K., Napel, S., Vorontsov, E., Maier-Hein, L., Cardoso, M.J., “A large annotated medical
-

- image dataset for the development and evaluation of segmentation algorithms”, 2019.
80. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017b. “ChestX-Ray8: Hospital-scale chest X-Ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases”, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 3462–3471.
  81. Kavur, A.E., Selver, M.A., Dicle, O., Barış, M., Gezer, N.S., 2019. “CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data”.
  82. Girshick, “R. Fast r-cnn”. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
  83. Ren, S.; He, K.; Girshick, R.; Sun, J. “Faster r-cnn: Towards real-time object detection with region proposal networks”. IEEE Trans. Pattern Anal. Mach. Intell. 2016, 39, 1137–1149.
  84. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
  85. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, “A. You only look once: Unified, real-time object detection”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
  86. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. “Ssd: Single shot multibox detector”. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
  87. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. “Rich feature hierarchies for accurate object detection and semantic segmentation”. arXiv, 2014; arXiv:1311.2524.

88. He, K.; Zhang, X.; Ren, S.; Sun, J. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition". arXiv, 2014; arXiv:1406.4729.
89. Wang, X.; Shrivastava, A.; Gupta, A. "A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection". arXiv, 2017; arXiv:1704.03414.
90. Ren, S.; He, K.; Girshick, R.; Sun, J. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". IEEE Trans. Pattern Anal. Mach. Intell. **2017**, 39, 1137–1149.
91. Joseph, R.; Santosh, D.; Ross, G.; Ali, "F. You Only Look Once: Unified, Real-Time Object Detection". arXiv, 2015; arXiv:1506.02640.
92. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. "SSD: Single shot multibox detector". arXiv, 2016; arXiv:1512.02325.
93. Simonyan, K.; Zisserman, "A. Very Deep Convolutional Networks for Large-Scale Image Recognition". arXiv, 2014; arXiv:1409.1556.
94. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, "Inception-ResNet and the Impact of Residual Connections on Learning". arXiv, 2016; arXiv:1602.07261.
95. Abacha, Asma Ben and Gayen, Soumya and Lau, Jason J and Rajaraman, Sivaramakrishnan and Demner-Fushman, Dina. "NLM at ImageCLEF 2018 Visual Question Answering in the Medical Domain" (2018).
96. Hochreiter, S., Schmidhuber, J., 1997. "Long short-term memory". Neural Computation 9, 1735–1780.
97. Jiang, M., Chen, S., Yang, J., Zhao, Q., 2020. "Fantastic answers and where to find them: Immersive question-directed visual attention", in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 2977–2986.
98. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S., 2019. "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs". arXiv preprint arXiv:1901.07042 .

99. Jung, B., Gu, L., HaradaAl-Sadi, T., 2020. bumjun\_jung at Healthcare Image QA 2020: “VQA model based on feature extraction and multi-modal feature fusion”, in: CLEF 2020 Working Notes.
100. K. Verma, H., Ramachandran S., S., 2020. “HARENDRAKV at Healthcare Image QA 2020: Sequential VQA with attention for medical visual question answering”, in: CLEF 2020 Working Notes.
101. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D. & Parikh, D. “Making the v in vqa matter: Elevating the role of image understanding in visual question answering”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6904–6913 (2017).
102. Allaouzi, I. & Ahmed, M. B. “Deep neural networks and decision tree classifier for visual question answering in the medical domain”. In CLEF (Working Notes) (2018)
103. C. Wang, H. Yang, and C. Meinel, “Image captioning with deep bidirectional LSTMs and multi-task learning,” *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 14, no. 2, 2018.
104. T. Liu, S. Yu, B. Xu, and H. Yin, “Recurrent networks with attention and convolutional networks for sentence representation and classification,” *Applied Intelligence*, vol. 48, no. 10, pp. 3797–3806, 2018.
105. Y. Yang, W.-T. Yih, and M. C. Wikiqa, “A challenge dataset for open-domain question answering,” in Proceedings of the Conference Empirical Methods Natural Language Processing, Lisbon, Portugal, September 2015.
106. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
107. B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, “Simple baseline for visual question answering,” 2015, arXiv:1512.02167
108. T.-Y. Lin, A. RoyChowdhury, and S. Maji, “Bilinear CNN models for finegrained visual recognition,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1449–1457.
109. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2001, pp. 311–318.

110. Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. 2015. “ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering”. arXiv e-prints (Nov. 2015), arXiv:1511.05960

# PUBLICATIONS



# Visual Question Answering using Data Mining Techniques for Skeletal Scintigraphy in medical domain - VQADMSS

*Mr. Jinesh Melvin YI*  
 Department of Computer Engineering  
 Pillai College of Engineering,  
 New Panvel, Navi Mumbai, India  
[yijmelvin@mes.ac.in](mailto:yijmelvin@mes.ac.in)

*Dr. Sushopti Gawade*  
 Department of Computer Engineering  
 Pillai College of Engineering,  
 New Panvel, Navi Mumbai, India  
[sgawade@mes.ac.in](mailto:sgawade@mes.ac.in)

*Dr. Hemant Palivela*  
 Head of AI and ML  
 eClerx Services Ltd.  
 Mumbai, India  
[hemant.datascience@gmail.com](mailto:hemant.datascience@gmail.com)

**Abstract**—Understanding about the medical images of patients is a very tedious task. Doctors should convey their patient through the image of the questions asked by the patient. Large amounts of labeled data are required for training in traditional approaches for VQA (Visual Question Answering). Also, the description of clinic trial text in English and in multilingual contexts is one of the challenges in the medical field. To present the clarification about the images, doctors are required to provide the related images. It is better for comparison with the patient's previous report and current report. This paper contributes to solve the problems related to VQA for better description of the image and accuracy related images through the answer of the questions, also to make it easy to convey the users with any kind of images. Question answer process should be more descriptive for easy to understand and traceable. This system helps to identify the types of images which are captured by any scanner. The better accuracy is a visualization method which projects the answers as a baseline that shows the corresponding region with various colors, which is easier to note the answers in visual method for the appropriate questions. This proposed framework focuses on Radiology image for Skeletal Scintigraphy to transform and generate a model using Data Mining Techniques. This system suggests that the effective medical Visual Question answering techniques is better to assist doctors in clinical analysis and diagnosis. This also will help the hospital services to grow the medical domain.

**Keywords**—Radiology images, Classification models, Generative models, Transformer, Visual Question Answering

## I. INTRODUCTION

Medical domain field is rapidly growing with different tools and techniques to improve the benefits of patients, researchers and Clinicians. Past few years, research in computer science and medical science have been developing intelligent tools for supporting medical decision making. Different software's was designed by various vendors to help the doctors, patients and clinicians. Researchers are also keen to provide new techniques with the help of technology to help society. The difficulties faced by patients are to understand about their health and body conditions. To know exact information about body and health communication between Doctor and patient communication is

very important. While discussing health conditions, patients may or may not understand the terms which are used by doctors or clinicians. Various soft computing systems have been successfully developed in health care professionals to support patients about this. Many radiology centers are available around for patient's benefits with the latest techniques and tools. In case of radiology images, patients need to consult some doctors regarding their health. Consulting people will raise many questions about the radiology image to clarify their health conditions. Manually it is difficult to answer all the queries of patients, so this research proposes automatic answers about the queries asked by patients.

### A. Overview of Radiology image

Radiology is an imaging technology that is used to diagnose and treat disease and it is a branch of medicine. Radiology has two different areas, diagnostic radiology and interventional radiology. Diagnostic radiology is used to visualize the structure inside the body. The specialist will use integration of these images. It helps to show all the symptoms and also to monitor how well the body responds to the treatment patient is receiving for specific disease. It helps to visualize different illnesses, such as colon cancer, heart disease and breast cancer etc. The most common types of diagnostic radiology exams include CT (Computed Tomography) also known as CAT (Computerized Axial Tomography). It includes CT angiography, Fluoroscopy with upper GI, magnetic resonance imaging (MRI) scan and magnetic resonance angiography (MRA) scan, mammography, bone scan, thyroid scan, plain x-ray, PET images, PET scan, PET-CT, ultrasound. As the user enters the input image, this system will compare with the database and convey it to the user with proper information for further process.

Interventional Radiology imaging is helpful to doctors when inserting wires, catheters and other small instruments and tools into our body; it is a smaller incision cut. It often involved treating blockages, problems in veins, problems in uterus, back pain, liver problems and kidney problems. Proposed system helps to identify Interventional Radiology images. This

research deals with both types of radiology images related to different diseases.

### B. Question Answer Processing

The proposed system VQADMSS supports question answering mechanisms to help patients as shown in Fig 1. The huge amount of data should be trained for predicting an accurate output, once the raw data gets trained with large amounts of questions with subjective and descriptive answers. The large datasets have a large number of variables to compute the resource to process them. The effective amount of data reduction without losing any data is one of the biggest challenges in this system. The reduction of data helps to reduce the machine effort and generate the new model to increase the speed up the performance and generalize the process of learning to the machine. After the feature extraction process the quality datasets get stored in a new location and data mining techniques will help to classify with a random forest algorithm to predict the output answer for the question from the image will be accurate. Test the machine with an image and ask different questions, so that the data mining classification algorithm gets processed and predicts the proper output and it asks whether to predict the related image for further clarification of input images and questions.

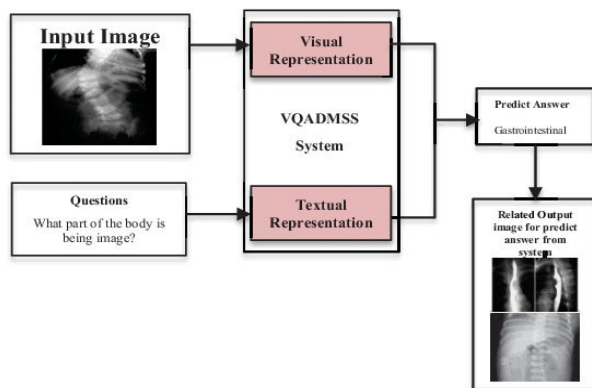


Fig. 1. Block diagram of QA System.

The various medical assists launched in the medical domain to improve and ease access, but Visual Question Answering (VQA) is somewhat different techniques also it gives benefits to any kind of patient. There are lots of ways to study about our health conditions without any expert guidance. Computer techniques played major roles in the medical domain and it grew in medical health and hospital services. The system which helps the patients will give more clarity in any type of radiology image using VQA. Our paper focused on Skeletal Scintigraphy, about bone marrow, bone cancer, density, infection, osteonecrosis, osteoporosis etc. It assists for the same to patients, also it helps with a multilingual system for illiterate people. It gives more predicted values during the Question Answering session.

Many Different questions considered by any patients have not satisfying the methods to solve their comprehensive problem, in QA System diagnosis CGMVQA model generates

capabilities to turn the complex problem to simple problems, which includes supervised and question generation, with data augmentation on images and tokenization on texts. It minimizes the parameter of the multi-head self-attention transformer to cut the computational cost down, and also add different kinds of embedding together to deal with text [2]. PICO is a framework for Data Mining in clinical Trial Text which transforms for classification and question answering tasks. It is mainly focused on detecting and annotation of information about PICO elements. The characteristics of PICO is Population, Intervention, Comparator and Outcome. It allows creation with the aim to support systematic reviews of semi automation using NLP method. It contributes to the sentence prediction task, training the data from different tasks and integrating. Prepared the dataset using training and testing subsets that fit the SQuAD format and merged both the dataset in order to check the correct answer for PICO questions and flexibility of general purpose of answering for questions on the basis of SQuAD [1].

Closed-end and Open-end are the two important components in medical visual question answering via conditional reasoning. There are multiple questions arise from users by two types of questions mainly underlined. In closed-end tasks only the answers should be limited like yes or no but in open-end tasks the answers should be free texts. Image is given as the input and output will be the answers in two forms. Question condition reasoning module to guide the modulation of multimodal fusion features [3].

## II. RELATED WORK

(Deepak Gupta, Swati Suman and Asif Ekbal, 2010) designed a new system with three different techniques such as Question Segregation techniques, then the second stage of this system is to integrate the question segregation model with hierarchical deep multi-model neural network and with the impact of question segregation which compare the performance of hierarchical deep multimodal network with question segregation and without question segregation. SVM algorithm is used as a base classifier to extract feature vectors for question segregation. This system produces two different types of input model which are yes/no and others. Examined this system with RAD and CLEF18 datasets in [4].

In the Question Answer with Deep Learning a survey has an Automatic Question Answering system takes place in three different groups, such as deep neural network, dynamic memory network and rational networks. In this the datasets to be two different categories one is textual and other one is in visual. Finally it evaluates the matrix for information retrieval system and automatically text generation. Deep Learning in neural networks which predict a solution for a task and it concludes with previous experience. The system gets processed between two sets of data that concern input and output to solve a task. The two main common architectures used for survey, they are Convolution Neural Network (CNN) and Recurrent Neural Network (RNN). To solve the sequence tag is to be used by dynamic memory networks also for classification problems, sequence to sequence tasks and question answering tasks. To compute the relation without the need to be learned in which recurrent neural networks learn to capture sequential dependency and Convolution Neural Network (CNN) learn spatial dependencies.

As on Image CLEF 2020: Multimedia Retrieval in Lifelogging Medical, Nature and Internet application follows four different tasks in which the first task is all about life log, it cares videos, images and other sources about daily activities understanding, retrieval and summarization. Then the next to analyze the caption, predict the tuberculosis and answer the questions of medical images which mean by Medical tasks. The segmenting and labeling collection of coral image for 3D modeling is a coral task and the final task is about the web user interfaces to detect and recognize the problem which were addressed from hand drawn websites for generating automatic code [6].

Domain Specific Question Answering system is based on a knowledge graph which is also meant by causality knowledge graph and it indicates the process of question answer. This relationship redefined and extracted the knowledge graph of doctor's. Once the question answer pair generated from concept knowledge graph to train the machine learning models for retrieval. Concept Knowledge Graph is used to display the schema of data, which improves the completeness. Doctors are invited to set some rules to convert classes into triples. So that the experts have enough knowledge about the relationship of the concepts in the medical domain. Instance knowledge graphs are specific entities like disease or drugs etc. Hybrid method is used to retrieve the answers. The system tries two different models for every Natural Language Question. The traditional method is used by support vector machines and sequence to sequence deep learning models. In which SVM classifies for discrete and continuous variables, but sequence to sequence deals with text. Structure query based models intend to translate the Natural Language Questions submitted by users into SPARQL Query statements. The system will first execute the word segmentation using the jieba word segmentation tool. Medical dictionaries tools help to improve its reliability. The system converts questions into vectors and CNN model classifies the questions to find best-fit then it will execute the SPARQL query and return the results to the users. Question Answer presents subgraphs with corresponding Natural Language sentences. It makes the answer reasonable. It helps users to get more knowledge or information about his or her questions [7].

To extract the visual feature from CNN (Convolutional Neural Network) & Global Average Pooling strategy. It captured the medical images using training datasets. The BERT model plays a major role to encode the question which is raised with semantic features. It scores the accuracy of 0.624 also the image CLEF 2019 has selected as a trained image. Bidirectional Encoder Representation from Transformers has been released two size BERT<sub>BASE</sub> and BERT<sub>LARGE</sub>. BERT<sub>BASE</sub> has 12 layers in the Encoder stack but BERT<sub>LARGE</sub> has 24 layers in the Encoder stack. BERT is an Encoder stack of transformer architecture through the network, which uses self-attention. The different stages of these systems extract the feature from the imported image, parallelly it encoded the semantic question using the BERT model, then it forward to feature fuse with co-attention mechanism. It helps to avoid irrelevant information for different regions. Image CLEF 2019 visual question answering in the medical domain has used multi-modal factorized bilinear pooling. The performance of this model is very efficient and effective for VQA, because of the combination with multi model features. But Bilinear is outperform, also its traditional linear modes for VQA. It gives limited capacity and low performance [9]. The final model is to predict the answers with

the accuracy score of 0.624, also the perfect mismatch score for this system is 0.644[8].

Two different classification models used in the medical VQA framework named MAML and CDAE. This model is used to initialize the model weights for image feature extraction. MAML represents Model-Agnostic Meta-Learning, training the model with a dataset for MAML by manually reviewing more number of question and answer pairs. The images from the dataset get separated into three categories such as head, chest and abdomen, also it gets divided into subgroups based on question answer pairs corresponding to the images. From all of these it categorized into 9 classes. Unlabeled images are used to train CDAE and encoder by feeding them before input images which use Gaussian noise to corrupt [11]. After training both MAML and CDAE this system used trained weight MEVF image feature extraction components in VQA framework, then finetune the whole VQA model using a train set of VQA-RAD dataset, which makes genuine comparison to [12]. The VQA accuracy is computed as the percentage for open-ended and close-ended for both MAML and CDAE models from scratch and finetuning. From this analysis finetuning has more accuracy than scratch.

### III. PROPOSED SYSTEM

In Visual Question Answering for Skeletal Scintigraphy system collect any type of Radiology image there is no limits of image selection. Diagnostic and Interventional are the two different types of Radiology image. It's an imaging technology to diagnose and treat diseases. This system focuses on skeleton scintigraphy which is considered as the study of bone scans. It helps to detect the fractures in bones, cancers in bones, insufficiency fractures, affection of bone joints and many others. This helps to answer the Question for all types of bones in the human body from skull to foot, such as Long bone, Short bone and irregular bones. The most common types of diagnostic radiology are computed tomography which we called CT scan also it named as Computerized axial tomography (CAT) scan, the other type of radiology images are magnetic resonance imaging (MRI) and magnetic resonance angiography (MRA), plain X-ray, positron emission tomography also known as PET imaging, PET scan or PET-CT and ultrasound. The other type of Radiology image is Interventional radiology, it helps the doctors to insert catheters, wires, tools and some small instruments into the human body. Doctors can detect and treat the diseases directly through a scope which is mentioned as a camera with open surgery, some of the Interventions are Nuclear Medicine Imaging, PET, X-ray, and Ultrasound.

Question answering process should be more descriptionable for easy to understand and traceable. This helps any kind of patient and doctors to get a clear view about the images, also it clear the maximum doubts by giving proper description. This system helps to identify the types of images captured by the medical imaginary tools like whether the image is Diagnostic Radiology or Interventional Radiology. Visualization the answers as baseline and it shows the corresponding regions based on the question. Suppose the doctors or patients want to know the explanation of the particular area from the image, so the system identifies the proper region which enquiry by the users by mapping the baseline as in visualization mode as mentioned in Fig. 3. As the description of answers predicted from the images

which the questions asked by the users as mentioned in Table 1, also this system helps to retrieve the related images from the answer as reference to the user, so that it is more understandable and convenient to the users for further treatments. It seems like the answers from images and the reference images predicted from the answers.

#### IV. SYSTEM DESIGN AND METHODOLOGY

The classification technique is the better suggestion in Machine Learning to develop this system. The huge amount of radiology medical images to be collected from various sources as a dataset for implementing the system, in which we use a set of images for training and test the ratio for supervised learning is 70:30. The 70% of dataset will be used to train the machine with different answers from the image like type of images, Question types, Name of the image, image position and which part of the organ. Before training the dataset, the image extraction module will extract the feature image from the normal image. The Inception Resnet V2 model is used for image feature extraction as it is a type of Convolutional neural network, which trains more than millions of images from the Image Net database. Also it classifies the images into 'n' number of object categories. It gives good performance and decreases the training cost. This model enables connection shortcuts to speed up network training. Bidirectional layer input models run in two ways, one from the past to future and other vice-versa. It helps to extract the question features, with LSTM. Bi-LSTM generates the representation of questions [4]. Once the image gets trained with the above ratio, then it is classified using machine learning algorithms. Random Forest is the best choice in supervised learning for classification the trained dataset with more accurate results. It has the direct relationship between the number of trees in the forest and the results will be in an efficient manner. Compared with other algorithms like SVM, Regression, K mean, the accuracy of RF is higher. RF constructs a decision tree from the given dataset for every sample and it will predict the results from every decision tree, then the voting will perform for every predicted result. The final prediction result is the most voted tree from the dataset. The predicted output will be the answer for the question asked by the user. Finally it fetches the related part of the image from the backend for more effectiveness or it refers the image from the answer for further treatments.

TABLE I. Sample Questions and answer pairs formulated from a single image, more than one medical related question asked from a given image.

Questions	Objective Answers	Subjective Answer	Organ	Image Type
What does the CT scan show?	left atrium	A large filling defect in the left atrium.	Fetal Skull	Diagnostic
Where is the anterior fontanel?	Top		Fetal Skull	Diagnostic
Is it normal?	Yes		Fetal Skull	Diagnostic

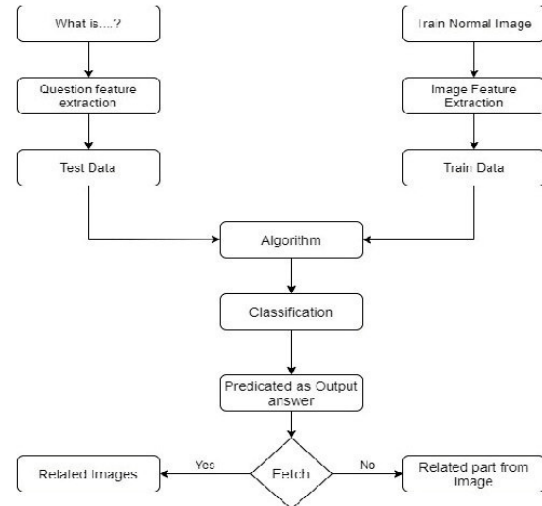


Fig. 2. Design of Proposed System.

#### V. USE CASE RELATED ON PROPOSED SYSTEM

The Organ shown in Fig 3 is Fetal Skull. It's an input image by the user, which has two types of view (i) Superior view and (ii) Lateral View. The result has Questions, Type of Questions, and Answers in Objective type and Subjective type, Name of the Organ and Image Type as shown in Table 1. The displayed image parts name will list out to the users in the same page as mentioned in Fig. 4, when the user clicks on any listed name it will show the part in the displayed image with base red line. The related image option will also be available on the same screen to get more clarity about the image.

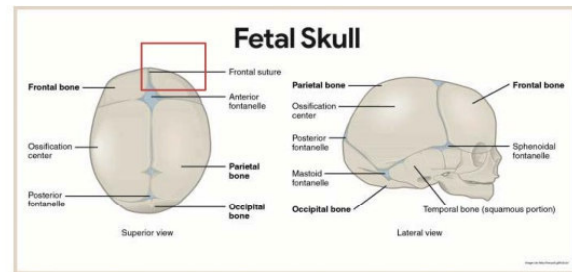


Fig. 3. Fetal Skull in Superior view and Lateral view [13]

Visualize Fetal Skull Names	
1.	Frontal Bone
2.	Ossification Center
3.	Posterior fontanels
4.	Frontal Suture
5.	Anterior
6.	Parietal bone

Fig. 4. List the names of Fetal Skull in Superior view

## CONCLUSION

This system will explore the Answering for users question in visual mode for medical images, it will help users to take further procedure after getting proper results. The contribution of this system is to train the normal images and extract the feature from the image and it trains the data for classification. It predicts the proper answers in a descriptive manner with better accuracy. Also it will retrieve the related images for user's reference. Different algorithms used in survey papers which have less accuracy also only identifies the answers for questions. The cost of training set is very low when we compare with other models as we prefer Resnet v2 model and Bi-LSTM for image and question feature extraction. Use cloud technology for more accessible to the users where they can utilize this system in any remote area with more predictable as future enhancement.

## REFERENCES

- [1] Lena Schmidt, Julie Weeds and Julian P. T. Higgins, "Data Mining in Clinical Trial Text: Transformers for Classification and Question Answering Tasks," University of Bristol, Bristol Medical School, 39 Whatley Road, BS82PS Bristol, UK, Jan 2020.
- [2] Fuji Ren, (Senior Member, IEEE), and Yangyang Zhou, "CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering," Faculty of Engineer, University of Tokushima, Tokushima 770-8506, Japan, March 6, 2020.
- [3] Li-Ming Zhan, Bo Liu, Lu Fan, "Medical Visual Question Answering via Conditional Reasoning", The Hong Kong Polytechnic University, October 2020.
- [4] Deepak Gupta, Swati Suman, Asif Ekbal, "Hierarchical deep multi-modal network for medical visual question answering", Department of Computer Science and Engineering, Indian Institute of Technology Patna, India, Sep 2020.
- [5] Martina Toshevska, Georgina Mirceva, Mile Jovanov, "Question Answering with Deep Learning: A Survey," Faculty of Computer Science and Engineering Ss. Cyril and Methodius University Skopje, Macedonia, 11 March 2020.
- [6] Bogdan Ionescu, Henning M'uller, Renaud P'eteri, "ImageCLEF 2020: Multimedia Retrieval in Lifelogging, Medical, Nature, and Internet Applications," University Politehnica of Bucharest, Bucharest, Romania, Aug 2020.
- [7] Ming Sheng, Anqi Li, Yuelin Bu, BNRist, "DSQA: A Domain Specific QA System for Smart Health Based on Knowledge Graph," DCST, RIIT, Tsinghua University, Beijing 100084, China, Aug 2020.
- [8] Xin Yan, Lin Li, Chulin Xie, Jun Xiao, and Lin Gu, "ImageCLEF 2019 Visual Question Answering in the Medical Domain," Zhejiang University, Hangzhou, China, Sep 2019.
- [9] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning M'uller, "VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019," Lister Hill Center, National Library of Medicine, USA, 2019 Sep.
- [10] Scibert: Pre-trained contextualized embeddings for scientific text. ArXiv, abs/1903.10676, Beltagy, L., Cohan, A., and Lo, K. (2019).
- [11] Jjj Binh D. Nguyen, Thanh-Toan Do, Binh X. Nguyen, Tuong Do, Erman Tjiputra, and Quang D. Tran, "Overcoming Data Limitation in Medical Visual Question Answering", AIOZ Pte Ltd, Singapore, 26 Sep 2019.
- [12] Lau, J.J., Gayen, S., Abacha, A.B., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Nature (2018)
- [13] <https://i.pinimg.com/originals/66/01/ee/6601ee3b13a9e24a6aa53942fe1f8ce1.jpg> is accessed on 20 Jan 2021.
- [14] Asma Ben Abacha, Soumya Gayen, Jason J. Lau, Sivaramakrishnan Rajaraman, and Dina Demner-Fushman. 2018. NLM at ImageCLEF 2018 Visual Question Answering in the Medical Domain. In Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum (CEUR Workshop Proceedings, Vol. 2125). CEUR- WS.org, Avignon, France.
- [15] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR. IEEE Computer Society, Salt Lake City, UT, USA, 6077–6086.
- [16] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," IEEE Access, vol. 6, pp. 9375–9389, 2018.
- [17] R. Wang, X. Liang, X. Zhu, and Y. Xie, "A feasibility of respiration prediction based on deep bi-LSTM for real-time tumor tracking," IEEE Access, vol. 6, pp. 51262–51268, 2018.

# Novel approach to integrate various feature extraction techniques for the Visual Question Answering System with skeletal images in the healthcare sector

**Jinesh Melvin Y I**

Computer Engineering  
Pacific Academy of Higher Education and Research University  
Udaipur, India  
jm3998@gmail.com

**Sushopti Gawade**

Computer Engineering  
Mumbai University  
Mumbai, India  
[sushoptikrishimitra@gmail.com](mailto:sushoptikrishimitra@gmail.com)

**Mukesh Shrimali**

Computer Engineering  
Pacific Academy of Higher Education and Research University, Pacific Hills  
Udaipur, India  
Mukesh\_shrimali@yahoo.com

**Abstract**— In the realm of medical science, one of the most challenging concepts to grasp is the Medical Imaging Query Response System. The comprehension and classification of the diverse representations of the human body require a significant degree of effort and expertise. Furthermore, it is imperative for users within the healthcare sector to rigorously validate the system. In the domain of human health, a plethora of imaging techniques, including MRI, CT, ultrasound, X-ray, PET-CT, and others, play a pivotal role in the identification of medical issues. These technologies are instrumental in supporting both patient engagement and clinical decision-making. However, the utilization of models, techniques, and datasets for processing textual and visual information introduces complexities that can at times impede the provision of pertinent clinical solutions. The overarching objective of the proposed approach is to conduct a comprehensive comparative analysis of various feature extraction methodologies for both visual and textual information within the Visual Question Answering (VQA) system, focusing on human skeletal images. This endeavor is aimed at enhancing the VQA system's performance with newer datasets and addressing any limitations inherent in existing models. In addition, this research initiative seeks to enable researchers to identify and optimize novel methods that enhance the accuracy of the VQA system. The models under scrutiny in this analysis encompass various methods of feature extraction that help to improve the model and quality of the healthcare industry. The researcher will find the proper methodology for different datasets. To gauge the efficacy of each model in delivering the desired outcomes, an array of metrics will be employed, including classification measurement accuracy, F-classification, C-true positive rate (CTPR), C-precision, C-recall, C-sensitivity, and C-false negative rate (FNR). These metrics are designed to enhance the accuracy of any dataset and optimize the performance of both visual and textual components to ensure accurate responses to the posed queries.

**Keywords**- Medical Images, VQA, Visual and Textual Feature Extraction methods, Classification model.

## I. INTRODUCTION

The field of medical science is experiencing rapid expansion, with a multitude of methods and strategies aimed at enhancing the welfare of patients, researchers, and clinicians alike. In recent years, the convergence of medical and computer science research has given rise to intelligent systems designed to facilitate medical decision-making. Diverse software solutions have been introduced by various providers to aid clinicians, patients, and healthcare practitioners. Researchers are enthusiastically embracing technology to pioneer novel approaches with the potential to benefit society.

Patients often grapple with comprehending the intricacies of their physical and medical conditions. In this context, the Visual Question Answering System has emerged as a prominent and invaluable research tool. This system finds its primary application in the realm of developing solutions capable of responding to queries based on visual imagery. The adoption of this technique has significantly bolstered decision-making processes across various domains and advanced technological applications.

The contemporary medical landscape is marked by swift expansion, encompassing the comprehensive scanning of the

human body through cutting-edge methodologies. While these scan datasets are predominantly in image format, manually deciphering the underlying textual context to address patient inquiries can be a daunting task. Within this context, our research focuses on medical image analysis, particularly within the domain of human skeletal imagery, leveraging an array of datasets available in the medical field.

The principal objective of this study is to identify and harness diverse datasets that facilitate the application of the Visual Question Answering (VQA) system. Additionally, this research seeks to assist medical professionals in making informed decisions while also providing valuable insights to researchers concerning system performance, thereby facilitating improvements through the development of new models catering to both visual and textual information.

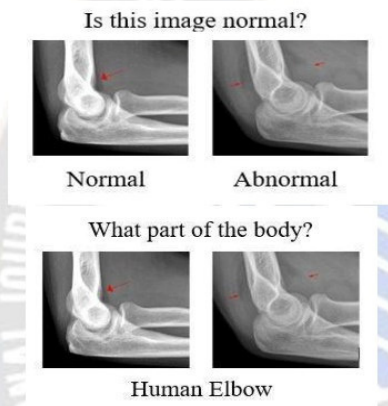


Figure 1. Actual Question Answer from Skeletal Image

In the realm of healthcare, there have been numerous advancements aimed at enhancing accessibility and facilitating medical assistance. Visual Question Answering (VQA) represents a unique approach that offers substantial benefits to a diverse range of patients. This method empowers individuals with the ability to conduct independent research on their medical conditions, reducing their dependency on healthcare professionals.

Over the years, computer technology has become increasingly prevalent within the healthcare sector, playing pivotal roles in various medical services. With the incorporation of VQA, patient-assistance systems are poised to significantly enhance the clarity and comprehension of diverse radiological image types.

Our proposed system is tailored to the specific domain of Skeletal Scintigraphy, encompassing a wide array of topics such as bone marrow, bone cancer, bone density, infections, osteonecrosis, osteoporosis, and more. This system not only assists patients in understanding these complex medical issues but also includes a multilingual feature to accommodate

individuals with limited literacy skills, ensuring inclusivity and accessibility for a diverse patient population.

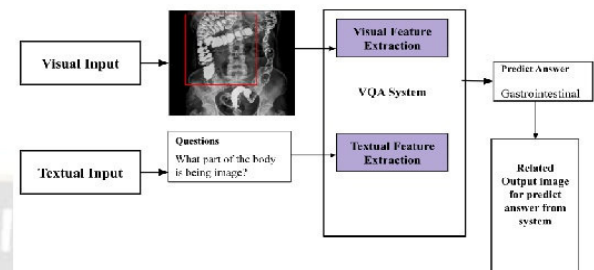


Figure 2. Visual Question Answer system for visual and textual input with Predictive answer

## II. PROBLEM STATEMENT FOR THE VQA SYSTEM

One of the difficult tasks in the medical industry is deriving useful information from medical imaging. The fundamental technology of question-answer systems is the extraction of precise user responses. Similar to the quickly expanding medical domain system, the input data extraction process needs to produce an effective and user-satisfying result. The most important component in classifying texts and images is feature extraction, which necessitates a deep understanding of the geometry and forms of real-world objects. Several classification methods entail performing data preprocessing operations, including normalization, identifying the classes, and extracting important features from the data cubes. In addition to making it easier for users to get images of any kind, the objective of solving VQA-related issues is to improve the description of the images and the accuracy of the related images by providing answers to the questions. For ease of understanding and traceability, the process of responding to the inquiry ought to be more descriptive.


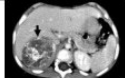
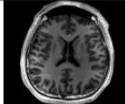
This technique aids in determining the kinds of images that every scanner captures. It is easiest to write below the answers in the visual technique for the corresponding questions when the visualization technique projects the answers as a baseline and displays the relevant region with numerous colors. This type of method yields the highest precision. In order to transform and construct a model utilizing classification, this suggested framework focuses on radiological imaging for bone scintigraphy. According to these ideas, the most useful medical methods for providing visual answers are those that help physicians with clinical analysis and diagnosis. Additionally, this will support hospital services in growing the medical field. Applications for classification techniques can be found in a variety of disciplines, such as traffic identification, medicine, and security. The textual and visual features can be extracted using the feature extraction model. For providing visual answers to questions about radiological imaging, it is the most effective approach. In this paper, we mentioned various datasets, various feature extraction methods, and their accuracy in the healthcare domain.


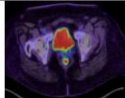
### III. RELATED WORK

There doesn't seem to be much research on VQA innovation that focuses on domains of professional healthcare yet. The visual mechanism of radiology images is complex, and the effect on visual perception is modest. Because people are anatomically similar, most radiological scans of the same body location in various people are rather comparable. Although the images seem to be the same, our inspection will reveal different problems. Furthermore, as Figure 1 illustrates, there are multiple questions that have been trained to model different responses. In [2], current datasets, sources, data quantities, and profession features were highlighted. Approaches were reviewed, suggestions were summarized, and future enhancements were planned. It encourages researchers working in this field. Figure 2 illustrates the structure of proposed design which has visual and textual inputs where medical or skeletal radiology images are considered as visual input and the questions related to input image are considered as textual input. The VQA system will extract the features of the input image using the Faster RCNN method and for text we preprocess the data, then extract the key features and finally integrate with classification methods for predicated output.

Radiology is a branch of medicine that uses imaging methods to identify, diagnose, and treat illnesses [8]. Two subspecialties of radiology are diagnostic radiology and interventional radiology [1]. Radiologists can assess internal body components using diagnostic radiography to seek out health problems, assess the source of symptoms, and track the body's response to treatment. The radiological modalities that are most frequently utilized are positron emission tomography (PET), magnetic resonance tomography, computed axial tomography, plain radiographic images, and ultrasound imaging [9]. It is helpful to visualize a variety of illnesses, including heart disease, colon cancer, and breast cancer. One of the most commonly used kinds of diagnostic radiology scans is CT (computerized tomography), also referred to as CAT (computerized axial tomography). Table 1 enumerates the diverse categories of radiological images alongside the corresponding medical terminology names for our system.

**TABLE 1** MEDICAL TERMINOLOGY FOR THE VQA SYSTEM

TYPES OF RADIOLOGY IMAGES	NAME OF IMAGE	IMAGES
X-ray Image	cervical spine	
CT Scan Image	Abdominal	
MRI Scan Image	Human cerebrum	

Ultrasound Image	Fetal	
PET Image	Sarcoma	

#### A. Challenges in Healthcare Datasets

Large-scale medical dataset preparation will require a great deal of work, and it should be done with due consideration for clinicians or physicians. Developing a medical VQA dataset is a highly challenging task. When creating a dataset, it is important to include photos from different radiology specialties, classify clinical questions for each image, have a solid understanding of medical terminology, and create precise responses for each question. We must lower the noise level of both the categorized question and answer because the noise level of the constructed dataset will be high. The dataset also includes a large number of photos with unclear pixels, objects, and other image errors. So it will be of absolutely no help for medical treatment, and it also includes questions that patients will find incomprehensible. Every image and response should follow the correct structure so that medical professionals can understand it. Another problem in the medical arena is scaling up the method to all unlabeled photos in the healthcare dataset.

#### B. Challenges in Feature Extraction Model

The existing Visual Question Answering (VQA) models employ Convolutional Neural Networks (CNN) to extract local regional vectors for specific areas within images. Long Short-Term Memory (LSTM) models are utilized to encode the feature vectors corresponding to the questions posed. While these models perform admirably in generating answers, they encounter limitations in scenarios where the response involves two adjacent local regions in the image and the question is structured as a complex sentence. It's worth noting that these models do not factor in the position and orientation of objects in their predictions.

Additionally, it's important to acknowledge that convolution operations are computationally more intensive and slower compared to max-pooling operations, both during forward and backward passes. Consequently, when dealing with deep networks, each training iteration naturally demands a substantially longer duration.

CNN-based algorithms necessitate extensive datasets to produce meaningful results, a limitation that can be challenging when dealing with scenarios involving a limited number of training instances. This is particularly significant considering the considerable resources, including time and expertise, required to compile and accurately categorize a comprehensive collection of images. In such cases, techniques like "data augmentation" and "transfer learning" are employed to address these limitations. Effective categorization heavily depends on the correct selection of image properties, as even the most

advanced machine-learning classifiers may perform poorly if these attributes are not appropriately chosen.

In addressing the vanishing gradient problem, Long Short-Term Memory (LSTM) models represent a noteworthy improvement over traditional Recurrent Neural Networks (RNNs). They expand the memory of RNNs to capture and retain long-term input dependencies. The "gated" cell within LSTM models empowers them to read, write, and erase information from memory, making informed decisions about which information to preserve or disregard.

The BiLSTM-CNN model employs Bidirectional LSTMs to encode both past and future contexts at each time step, following the CNN's encoding of each word. While this is beneficial for tasks like machine translation and sentence classification, it poses limitations for sequence-labeling tasks such as Named Entity Recognition (NER), as each token utilizes its own midway hidden states, unable to bridge past and future context effectively.

This research encompassed diverse datasets, various image feature extraction models, and textual feature extraction models, with summarized results presented in the following table.

**TABLE 2** INSIGHTS FROM THE LITERATURE SURVEY WITH VARIOUS DATASETS, FEATURE EXTRACTION AND ITS ACCURACY

MODEL	DATASETS	IMAGE FEATURE EXTRACTION	TEXT FE	CLASSIFICATION	CATEGORIES OF QUESTION
Vision-Language Model	VQA-RAD and PathVQA	ViT32 Model	BERT	Contrastive language-image pre-training (CLIP) model	Open-ended, Closed-ended
BPMVQA	VQA-Med 2018, Image CLEF 2019, VQA-RAD	CNN model to extract the spatial features	PubMed	Self-attention module and a feed forward neural network (FFN)	What, where, Yes/No
MedFusionNet	Image CLEF 2019	CNN models	BERT	MFB	modality

					Plane Organ
Adversarial VQA benchmark	Human-And-Model-in-the-Loop Enabled Training (HAMLET)	Adversarial VQA	SOTA models	-	Counting, OCR, Reasoning, Visual concept recognition

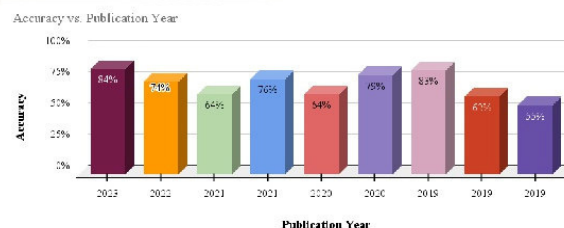


Figure 3. Existing Accuracy based on Publication Year

#### IV. METHODOLOGY

The system records any kind of radiological image; there are no restrictions on the kind of image that can be chosen when responding visually to questions related to bone scintigraphy. Diagnostic and interventional radiology images are two distinct categories. It is an imaging technology that helps in illness diagnosis and treatment. The system is designed to examine the skeletal scan, which is the equivalent of the bone scan aids in the detection of numerous conditions, including bone joint disorders, insufficiency fractures, shattered bones, and bone cancer. This provides an answer to the issue for every kind of bone in the human body, including long, short, and irregular bones, from the head to the foot. To make the process of asking and answering questions easier to comprehend, there should be more description in the process. This makes it easier for all patients and physicians to view the images clearly and eliminates the majority of doubts with a thorough description. This system serves the purpose of categorizing images produced by medical imaging tools, distinguishing between images from diagnostic radiology and interventional radiology. It enables users to pose questions related to these regions. As described in Figure 1, it provides answers based on user-generated questions and retrieves pertinent images in response. This feature greatly enhances user understanding and facilitates follow-up care. The referenced images inferred from the answers are consistent with the image-based responses.

##### A. Use case related to Proposed System

The fetal skull organ, as illustrated in Figure 3, is the subject of discussion. In this context, the system enables users to input an image featuring two distinct perspectives: the superior view and

the lateral view. The system furnishes the results in Table 3, presenting details such as the questions posed, the types of questions (both objective and subjective), the identified organ, and the image type, as detailed in Table 2. Upon selecting any of the organ names listed in the table, users will be directed to the section of the displayed image marked by a red line at its base. Furthermore, the relevant organ names associated with the displayed image segment will be presented to users on the same page, in line with the depiction in Figure 3. Users will also have the option to access additional information regarding the image via the related image selection, conveniently available on the same screen. This approach aims to offer users a seamless and informative experience.

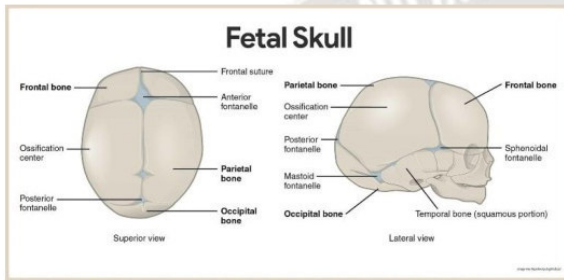


Figure 4. Fetal Skull in Superior view and Lateral view

The collection contains about 3000 radiology and skeletal images from the medical domain. The training, testing, and cross validation ratios are 70:30. 70% of images from the preprocessed dataset will be used for training, while the remaining 30% will be used for testing and cross validation. In the same ratio, questions and answers will be trained from preprocessed images.

**TABLE 3** FORMULATED SINGLE IMAGE WITH VARIOUS QUESTIONS

Questions	Objective Answers	Subjective Answers	Organ	Image Type
What does the CT scan show?	left atrium	A large filling defect in the left atrium.	Fetal Skull	Diagnostic
Where is the anterior fontanel?	Top		Fetal Skull	Diagnostic

Is it normal?	Yes		Fetal Skull	Diagnostic
---------------	-----	--	-------------	------------

#### 1) Visual and Textual Feature Extraction

Most cutting-edge medical VQA systems rely on deep learning methods like attention mechanisms and recurrent neural networks (RNNs) [12] for text embedding and feature extraction, and convolutional neural networks (CNNs) for visual feature extraction. Deep learning transformers have been developed and successfully used for the medical VQA requirement. Transformers, for example, were originally applied to NLP applications like speech recognition [14] and machine translation [13]. The self-attention mechanism is the only source of dependency for its encoder-decoder design. Transformers show promise in learning relationships among sequence elements, in contrast to RNNs, which process sequence items recursively and only consider immediate context. Transformer designs that focus on entire sequences have the potential to learn long-range correlations. Specifically, the most commonly used model for textual information encoding is the bidirectional encoder representation from transformers (BERT) [15]. Using large-scale unsupervised corpora and a bidirectional attention mechanism, the language model BERT generates a context-sensitive representation for every word in a sentence. R-CNN's limitations were addressed with the introduction of Fast R-CNN. To create a convolutional feature map in this case, we simply send the input to CNN. From there, we identify the region proposals and use the ROI pooling layer to warp them into squares. The size can be changed, and it can feed into layers that are completely connected. It feeds none of the 2000 areas. According to the image, it immediately created the feature map. Compared to RCC, it is much faster for testing and training.

A faster R-CNN, also known as Fast R-CNN, is employed to identify region suggestions for selective search. It increases training and testing speed. The time it takes to get the output is decreased. To find the region proposals, a convolutional feature map performs better than a selective search technique.

The deep belief network training algorithm DBN can be used to initialize the network with random weights. Next, unsupervised learning can be used to train each layer of the network, starting from the first layer and continuing through the last layer. Finally, supervised learning and backpropagation can be used to fine-tune the entire network. This process must be repeated until the network has converged.

BiLSTM, or Bidirectional Long Short-Term Memory, comprises two separate LSTM neural networks, each with its own unique set of weights and bias factors. The outputs from the hidden layers of the forward and backward networks are combined through concatenation to form the feature vector that is subsequently extracted. In a study conducted by Linqin Cai, Sitong Zhou, Xun Yan, and Rongdi Yuan in 2019, they extensively discuss the operation of the Stacked Bidirectional Long Short-Term Memory Neural Network (SBiLSTMNN). They also delve into the coattention mechanism for question

representation and the attentive attention mechanism for answer representation. This comprehensive approach aims to provide a deep understanding of the SBiLSTMNN and its associated mechanisms.

## 2) Analysis of Experimental Results

Various datasets were analyzed from various research papers, which are mentioned in Table 2. Different categories of data that were used, its question answer type, and images, as shown in Fig. 4. It describes the total amount of data that leads to the ratio needed to train the model.

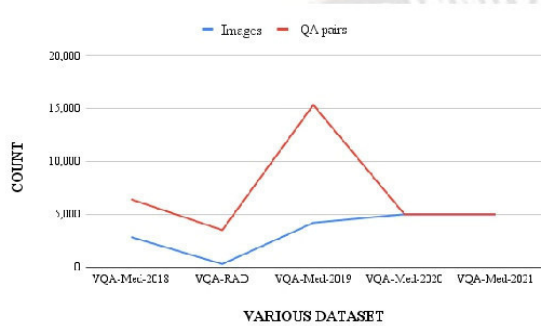


Figure 5. Types of datasets with total count of images and text mentioned in Table 2

Figure 5 depicts the total number of questions and images available to trained models. Each image has numerous questions, each in its own category.

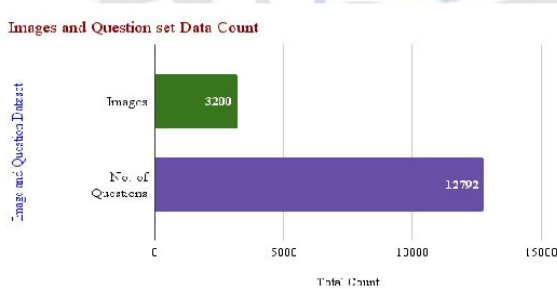


Figure 6. Visual and Textual Datasets

The many sorts of questions are depicted in the image below; each includes over 3000 question and answer sets to train the model, which is sufficient to develop the system. This helps to categorize the question and makes it easier to find related responses to the user's question. This type of question and answer was employed in the majority of previous models. Figure 6 illustrates the cumulative count of questions within each dataset category, each encompassing more than 3000 questions.

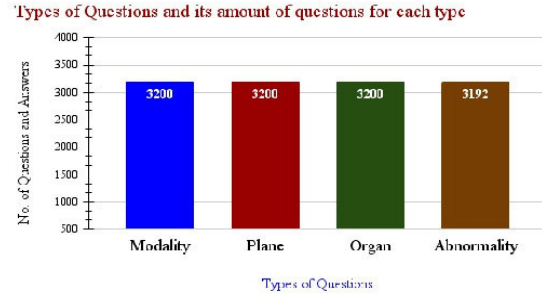


Figure 7. Category of Questions from dataset with a Total Count for each type

## B. Performance Metrics

The assessment of each existing model's performance involves the consideration of several key metrics, such as mean absolute error, mean squared error, root mean squared error, false measure, and precision.

Mean squared error (MSE) is a crucial metric that enables us to determine the average of the squared differences between the ground truth value ( $Y_j$ ) and the predicted regression value ( $Y'$ ).  $N$  represents the number of data points, as per equation (1).

In contrast, the mean absolute error (MAE) calculates the average of the differences between the ground truth and the predicted values, providing insights into the extent of deviations between forecasts and actual outcomes. It's worth noting that MAE employs the absolute value of the residual, making it direction-agnostic, meaning it doesn't discern whether under- or over-prediction has occurred. As outlined in equation (2), MAE is particularly robust against the influence of outliers.

This formal evaluation methodology aims to rigorously assess and compare the performance of these models in a quantifiable manner.

$$MAE = \frac{1}{N} \sum_{j=1}^N (Y_j - Y'_j)^2 \quad (1)$$

$$MSE = \frac{1}{N} \sum_{j=1}^N |Y_j - Y'_j| \quad (2)$$

The root mean squared error (RMSE) plays a significant role in the assessment of model performance. It calculates the average of the squared differences between the target value and the value predicted by the regression model. RMSE is particularly valuable because it rectifies a potential limitation of MSE, which excessively penalizes smaller errors by taking the square root of the result.

This square root transformation ensures that the scale of error interpretation aligns with that of the random variable, simplifying the process of understanding and analyzing errors. Essentially, RMSE normalizes the variables, reducing the potential impact of outliers on the overall analysis. This normalization is exemplified in equation (3).

In a formal evaluation context, RMSE provides an effective means of assessing the models, taking into account the scale of errors, and facilitating their meaningful interpretation.

$$RMSE = \sqrt{\frac{1}{N} \sum_{j=1}^N (Y_j - Y'_j)^2} \quad (3)$$

The Fmeasure range of feasible feature extraction approaches for both visual and textual datasets is depicted in Figure 7. For our datasets, the basic CNN has a lower level of Fmeasure than RNN and DBN.

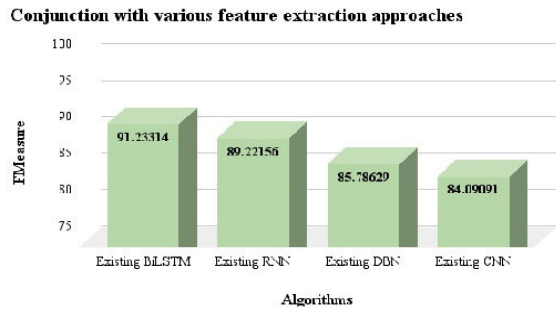


Figure 8. Conjunction with various feature extraction models

### C. Significance of the study

The exploratory outcome of the content extraction study, as shown in Figure 8, uses the removal to calculate the degree of coordination between the inquiry vector and the reaction vector. Manjunath Jogin and Mohana, May 2018, investigation study for execution of various categorization computations in Table 3. Consider the present models that have low accuracy for image highlight extraction and question reply feature extraction in Jinesh Melvin Y I, Sushopti Gawade, and Hemant Palivela, May 2021. The goal of the same paper was to describe the Visual Address Replying Framework for Radiology Images from Human Skeletal.

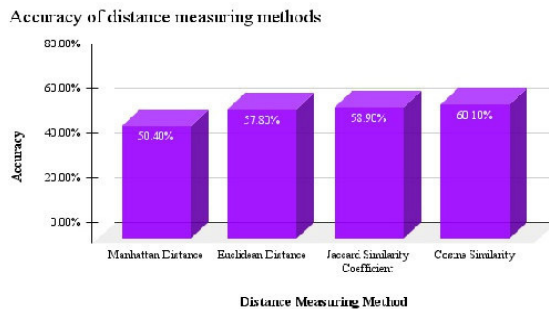


Figure 9. Accuracy with its distance measuring method for classifying various datasets

## V. CONCLUSION

A comparative analysis of diverse feature extraction methodologies was conducted, employing a range of distinct

datasets. This analytical approach serves to provide valuable insights for researchers striving to enhance the healthcare system's diagnostic capabilities for patient health assessment. Datasets were meticulously collected from a multitude of sources, facilitating a comprehensive evaluation of the existing methodologies within question-answering systems. The intended outcome of this endeavor is to contribute to the advancement of healthcare, ultimately enhancing the efficiency and effectiveness of patient outcomes. The future adoption and utilization of medical Visual Question Answering (VQA) systems will be contingent upon several pivotal factors. These factors encompass the abundance and caliber of medical VQA datasets, the development and evaluation of medical VQA models, as well as the seamless integration and practical deployment of medical VQA systems within clinical contexts. A critical imperative involves the generation of expansive, comprehensive, and heterogeneous medical VQA datasets that encompass a diverse spectrum of modalities, medical conditions, question types, and corresponding responses.

## VI. DECLARATIONS

### A. Funding

The authors specifically state that they received no financial aid, grants, or other forms of assistance to facilitate their research. This declaration emphasizes the research's independence and lack of outside influences on its findings.

### B. Statement on Conflicts of Interest

This work's authors have reported no conflicts of interest connected to the subject matter.

### C. Ethics Declaration

The author explicitly declares a lack of awareness regarding any personal or professional conflicts that might have influenced the research presented in this study. This statement underscores the commitment to maintaining impartiality and objectivity in the research.

### D. Code and Data Availability Statement

We used data from a variety of publicly available sources for the research, such as medical visual question answers from CLEF. This allows us to evaluate a variety of existing models and create a new framework for our system. The custom code is used to develop the application, which is used by us. The code for this project is confidential.

## AUTHORS CONTRIBUTION STATEMENT

Jinesh Melvin Y. I. is the corresponding author for the said manuscript. Jinesh Melvin Y.I. and Sushopti Gawade conceived of the presented idea. Jinesh Melvin Y. I. developed the theory and performed the computations. Sushopti Gawade and Mukesh Shrimali verified the analytical methods, encouraged Jinesh Melvin Y I to investigate the proposed work, and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

Jinesh Melvin Y I developed the theoretical formalism, performed the analytic calculations, and performed the numerical simulations. Both Jinesh Melvin Y I, Sushopti

Gawade, and Mukesh Shrimali contributed to the final version of the manuscript. Sushopti Gawade and Mukesh Shrimali supervised the project.

# REFERENCES

- [1] Y. I. Jinesh Melvin, Sushopti Gawade, Hemant Palivela, "Feature Extraction from Radiology Images for Visual Question Answering System Using CNN and BiLSTM Model", *Recent Innovations in Computing*, vol.832, pp.317, 2022.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [2] Yakoub Bazil, Mohamad Mahmoud Al Rahhal 2, Laila Bashmal 1 and Mansour Zuair 1 "Vision–Language Model for Visual Question Answering in Medical Imagery", *Bioengineering* 2023.
- [3] Stefania Barburiceanu, Serban Meza, Bogdan Orza, Raul Malutan, Romulus Terebes."Convolutional Neural Networks for Texture Feature Extraction. Applications to Leaf Disease Classification in Precision Agriculture", *IEEE Access*, 2021.
- [4] Y. Lu and S. Young, "A survey of public datasets for computer vision tasks in precision agriculture", *Comput. Electron. Agricult.*, vol. 178, Nov. 2020.
- [5] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning", *J. Big Data*, vol. 6, no. 1, pp. 60, 2019.
- [6] N. Ganatra and A. Patel, "A survey on disease detection and classification of agriculture products using image processing and machine learning", *Int. J. Comput. Appl.*, vol. 180, no. 13, pp. 7-12, Jan. 2018.
- [7] M. D. Zeiler, R. Fergus, "Visualizing and understanding convolutional networks", *ECCV*, 2014.
- [8] Herring W, *Learning radiology: Recognizing the basics*. Elsevier Health Sciences, 2015.
- [9] Novelline RA and Squire LF, *Squire's fundamentals of radiology*. La Editorial, UPR, 2004.
- [10] N. Ganatra and A. Patel, "A survey on disease detection and classification of agriculture products using image processing and machine learning", *Int. J. Comput. Appl.*, vol. 180, no. 13, pp. 7-12, Jan. 2018.
- [11] Sima Siami-Namini, Neda Tavakoli, Akbar Siami Namin, "The Performance of LSTM and BiLSTM in Forecasting Time Series ", *IEEE International Conference on Big Data (Big Data)* 2019.
- [12] Mikolov, T.; Karafiat, M.; Burget, L.; Cernocky, J.; Khudanpur, S. Recurrent Neural Network Based Language Model. *Interspeech* 2010, 2, 1045–1048.
- [13] Wang, Q.; Li, B.; Xiao, T.; Zhu, J.; Li, C.; Wong, D.F.; Chao, L.S. Learning Deep Transformer Models for Machine Translation. *arXiv* 2019, arXiv:1906.01787.
- [14] Chen, N.; Watanabe, S.; Villalba, J.A.; Zelasko, P.; Dehak, N. Non-Autoregressive Transformer for Speech Recognition. *IEEE Signal Process. Lett.* 2021, 28, 121–125.
- [15] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* 2019, arXiv:1810.04805.
- [16] Xin Yan, Lin Li, Chulin Xie, Jun Xiao, and Lin Gu, "ImageCLEF 2019 Visual Question Answering in the Medical Domain," *Zhejiang University, Hangzhou, China*, Sep 2019.
- [17] Lubna A, Saidalavi Kalady, Lijiya A., "MoBVQA: A Modality based Medical Image Visual Question Answering System", 978-1-7281-1895-6/19/\$31.00 c 2019 IEEE, 2019 IEEE Region 10 Conference (TENCON 2019).
- [18] Asma Ben Abacha, Soumya Gayen, Jason J Lau, Sivaramakrishnan Rajaraman, and Dina Demner-Fushman, "NLM at ImageCLEF 2018 Visual Question Answering in the Medical Domain", *CEUR-WS.org/Vol 2125/paper\_165.pdf*, Conference Paper · October 2018
- [19] Manjunath Jogin, Mohana, Madhulika M S, Divya G D, Meghana R K, Apoorva S, "Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning", 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT-2018), MAY 18th & 19th 2018.
- [20] Zhou Yu, Jun Yu, Jianping Fan, Dacheng Tao, "Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering", *arXiv:1708.01471v1 [cs.CV]* 4 Aug 2017

# CERTIFICATES



## CERTIFICATE OF PRESENTATION

*This certificate is awarded to*

**Mr. Jinesh Melvin Y I**

*for successfully presenting a paper at the*


**International Conference on Artificial Intelligence and Smart Systems (ICAIS 2021)**  
**organised by JCT College of Engineering and Technology, Coimbatore, India**  
**on 25-27, March 2021.**

**Paper Title: Visual Question Answering using Data Mining Techniques for Skeletal Scintigraphy in medical domain - VQADMSS**

**Author/s: Mr. Jinesh Melvin Y I, Dr. Sushopti Gawade, Dr. Hemant Palivela**

  
Session Chair

  
Conference Chair  
Dr. K. Geetha

  
Principal  
Dr. V. J. Arulkarthick



**1<sup>st</sup> Springer CCIS International Conference on  
Role of AI in Bio-Medical Translations' Research for the Health Care Industry**

**Organized by**

**G H RAISONI COLLEGE OF ENGINEERING, NAGPUR**

**Certificate**

to

**Jinesh Melvin Y I, Sushopti Gawade**

**Visual Question Answering System for Skeletal Image based on  
feature Extraction using Faster RCNN AND Kai-Bi-LSTM Techniques.**

**Paper Title:**

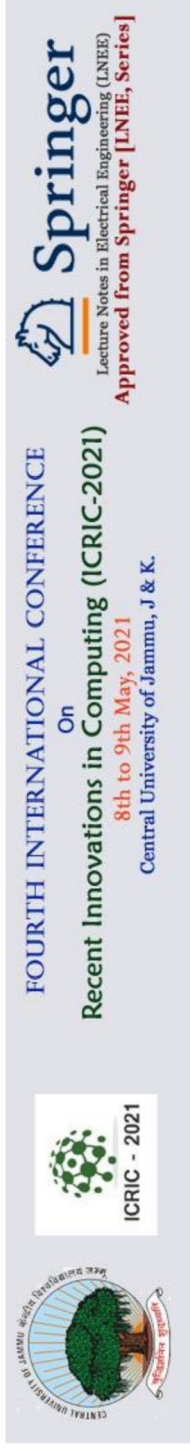
has presented in 1<sup>st</sup> Springer CCIS International Conference on “**Role of AI in Bio-Medical Translations' Research for the Health Care Industry**”(AIBTR-23) held on 23<sup>rd</sup> September, 2023 organized by Center of Excellence Biomedical Engineering & Technology incubation Center (BETiC-GHRCE) and Artificial Intelligence & Machine Learning (AIML) of G H Raisoni College of Engineering, Nagpur (India).

**Dr. P. Sivaram**  
General Chair

**Dr. Vikas Bora**  
General Chair

**Dr. Sachin Untawale**  
Honorary Chair & Director, GHRCE





## Fourth International Conference on Recent Innovation in Computing

ICRIC-2021

May 08-09, 2021

Organized by

**Department of Computer Science & Information Technology**

**CENTRAL UNIVERSITY OF JAMMU, J&K, INDIA**

***Certificate of Presentation/Participation***

This is to certify that Mr./Ms. **Jinesh Melvin Y I** of **Pillai College of Engineering** participated/presented a paper entitled **Feature Extraction from Radiology images for Visual Question Answering system using CNN and BiLSTM model**. during 4<sup>th</sup> International Conference on Recent Innovation in Computing organised by Central University of Jammu, Jammu, J & K, India.

**Dr. Yashwant Singh**  
**General Chair**

