International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(3s) 2026 pp. 103-111
https://doi.org/10.70917/ijcisim-2026-2282
Article

AI-Based Video Question Answering for Healthcare Applications Using Relevant Segment Localization

Indrajeet Kumar1*, Anand Kumar Gupta2

[1,2]Associate Professor, School of Engineering, P P Savani University, Surat, Gujarat, India
Corresponding Author: Indrajeet Kumar (Indrajeet.kumar@ppsu.ac.in)

Abstract: One important finding in real-world Video QA assignments is that, rather than needing full-sequence analysis, the textual question typically relates to a specific, brief segment of the entire movie. This encourages a more focused and effective learning strategy. In order to overcome this difficulty, we pre-sent Locate Before Answering (LocAns), a novel end-to-end architecture that performs answer prediction using only the localized section after first identifying the most pertinent temporal segment matching to the question. The two main parts of the suggested LocAns model are a response prediction module and a question localization module, both of which are integrated into a single pipeline. One significant improvement in our approach is the generation of training supervision. LocAns cleverly uses the ground-truth response labels to produce pseudo temporal supervision rather than depending just on manually marked temporal boundaries, which are sometimes unavailable or costly to collect. Three benchmark datasets such as NExT-QA, ActivityNet-QA, and AGQA are designed for long-term VideoQA were used for extensive research. In all three datasets, LocAns regularly beats current state-of-the-art techniques. In addition to producing excellent quantitative results, the model per-forms well qualitatively, as demonstrated by case studies in which it correctly identifies the most pertinent video segments prior to producing the right response. The locate-before-answer paradigm's efficacy is further supported by the localization module's enhanced interpretability. Overall, this study emphasizes how crucial temporally focused reasoning is while responding to lengthy video questions. LocAns creates a potential path for future Video QA research, especially for large-scale, complicated, and real-world video scenarios, by eliminating redundancy and improving the alignment between the question and pertinent visual evidence.
Keywords: Locate Before Answering (LocAns), Multi-Modal Learning, NEXT-QA, AGQA, Video Segment Selection,

1. INTRODUCTION

Video QA is becoming a standard for assessing a model's capacity to simultaneously comprehend dynamic visual scenes and human language as research in this field expands. Traditional image-based Visual Question Answering (VQA) and Video QA are conceptually and technically different [1]- [5]. In VQA, the model analyzes a single image that has all the information needed in a single static frame. But Video QA adds a temporal component.

A video is made up of many frames, each of which captures a separate instant across time, as opposed to a single image. Because the model must comprehend the order of events, causal relationships, actions, and scene transitions, this temporal progression complicates the reasoning process. Because short films often have a single event, a consistent background, and little action diversity, early Video QA research mostly concentrated on short video clips, typically lasting 10 to 15 seconds. Previous Video QA techniques were able to handle the full video as a consistent block of data because of these simpler features. In order to produce a single video-level representation, many algorithms processed the entire sequence by combining features from every frame [6] – [18]. This approach frequently yielded satisfactory results since short films usually preserve a single scene and a continuous action flow.

The longer the video, the more severe the noise and redundancy issue becomes. Unrelated frames make optimization more challenging, diminish the model's attention, and lower its representational quality. Therefore, it is inefficient and frequently unproductive to simply transfer techniques meant for short movies to long-term Video QA.

First, developing a dependable supervisory signal that can direct the temporal attention process is very difficult [3]. The majority of datasets do not specify the precise frame range from which the solution should be deduced; instead, they just offer the final result as a label. As a result, the model makes an effort to ac-quire localization solely through indirect answer monitoring, which frequently leads to erratic or unclear attention patterns.

Second, rather than identifying a coherent section of activity, temporal attention distributions frequently highlight fragmented and discontinuous frames [4]. It is challenging to understand the model's behavior and determine whether it is paying attention to significant evidence because of this dispersed attention. Additionally, the discontinuity impairs the model's capacity to record developing events, which are crucial for long-term video comprehension.

This study suggests a novel paradigm known as "Locate Before Answering" to overcome these constraints [5 -20] [32] [36]. The main concept is intuitive and reflects how people handle comparable issues: when posed a question about a lengthy movie, a person will first pick the pertinent section of the video and then thoroughly scrutinize that piece to determine the answer. The absence of temporal localization annotations in current VideoQA datasets, such as NExT-QA, ActivityNet-QA, and AGQA, is a significant obstacle to implementing this approach [25 – 30]. These datasets only offer question-answer pairings; they do not identify the specific video clip that answers each question. Although analogous issues without explicit labels have been attempted to be solved using weakly supervised Video Tem-poral Grounding (Video TG) algorithms, they usually function as standalone tasks in a two-stage pipeline where the localization module and the prediction module are trained independently [6], [30] – [39].

However, there are issues with training stability because of this bidirectional dependency. While QL depends on AP's performance, AP also needs to provide feedback to QL. LocAns uses an alternating training technique to guarantee con-vergence. This method stabilizes learning and does away with the need to manually balance the losses of both modules.

2. RELATED WORKS

Over time, Video Question Answering (Video QA) has grown to be a significant research issue in the field of vision-language understanding. Image-based Visual Question Answering (VQA), which focuses on evaluating static images, is seen as a logical extension of this task [7]. Although VQA has contributed to the advancement of multimodal learning, Video QA adds more complexity since im-age-based models cannot manage the developing scenes, temporal interdependence, and sequential activities found in movies.

The architecture used in early Video QA projects was often somewhat simple. Initially, 2D convolutional neural networks (CNNs) in conjunction with recurrent neural networks (RNNs) or spatiotemporal models like 3D CNNs were used to extract video information. Lan-guage features were often acquired using contextual models like BERT or embedding techniques like GloVe [8]. A cross-modal reasoning module was used to align the data from both modalities and produce the solution after visual and textual features were removed. Video QA research focused considerably more on understanding how visual information changed across frames and modeling temporal relationships than Image QA methods.

Despite their contributions, the majority of these attention-based techniques were created and assessed using early VideoQA datasets, including TGIF-QA, MSRVTT-QA, and MSVD-QA [9]. The videos in these databases are usually brief, lasting no more than 10 to 15 seconds. Because short movies seldom include intricate scene transitions or unconnected visual occurrences, the process is made much easier by the short temporal period. A number of long-duration datasets have been released recently in response to the growing demand for more accurate VideoQA standards. These consist of TVQA, AGQA, ActivityNet-QA, and NExT-QA [10].

These benchmarks include minute-level movies that span several activities and have noticeably deeper temporal structures, in contrast to short-video datasets. Among these, TVQA is unique as a tri-modal dataset since it contains question-answer pairs, videos, and subtitles. TVQA is therefore especially focused on language-driven reasoning and conversation comprehension [11] [32 - 39]. However, TVQA is less appropriate for researching pure video-based reasoning, which is the subject of this study, due to its strong reliance on text ex-changes. Thus, the

three well-known datasets that focus on visual and temporal reasoning in complicated, untrimmed videos—NExT-QA, ActivityNet-QA, and AGQA—are used for our investigations.

The Multi-Instance Learning (MIL) framework is a key component of weakly supervised VideoTG techniques [12]. This system treats every video as a bag of possible video segments, and it uses ranking losses to teach the model to distinguish between positive and negative segment-query combinations. The Text-Guided Attention (TGA) model by Mithun et al. is a noteworthy early effort in this approach that uses text-driven attention processes to learn to highlight query-relevant frames. Cross-Sentence Mining (CRM), developed more recently by Huang et al., enhances temporal comprehension by examining the connections between various sentences and utilizing contextual linguistic structures [13]- [21].

Weakly supervised VideoTG and our work are philosophically comparable in that they both seek to localize textual inquiries in videos without using temporal boundaries that have been annotated [22]-[30]. The objectives and methodological strategies, however, are very different. In contrast to VideoTG, which compares and ranks candidate moments using MIL methods, our approach directly employs answer annotations from VideoQA datasets to direct localization [30]- [39]. We create pseudo labels based on how well various suggestions promote accurate answer prediction, rather than depending only on implicit representations. More focused and task-specific localization is made possible by the close interaction between the answering and localization modules.

3. METHODOLOGY

The goal of the Video Question Answering (Video QA) challenge is to choose the right response (A) from a predetermined list of potential responses given an un-trimmed video (V) and a natural language question (Q). Long-duration videos, which frequently last for several minutes, have a lot of irrelevant information, numerous scene changes, and numerous actions. As a result, current models often become confused and have poor reasoning when the entire video is used as input.

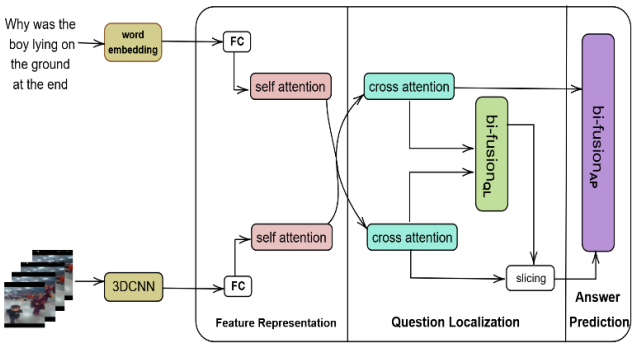

Fig. 1. The overall inference pipeline of the LocAns framework is shown here. The model is built around three key components: a feature representation module, a question localization module and an answer prediction module.

It's important to note that the majority of questions only touch on a tiny section of the film rather than its whole length. Based on this concept, the suggested approach, LocAns (Locate Before Answering), first determines which part of the video the question is referring to, then uses only that pertinent portion to deduce the response. By lowering noise and concentrating reasoning on the appropriate temporal region, this two-step method enhances performance.

The entire inference pipeline of the suggested LocAns framework is shown in Fig.1, emphasizing how the model analyzes the query and the video to get a precise response. In order to handle lengthy, untrimmed films, the pipeline first determines which portion is most pertinent before attempting to respond to the query.

First, two distinct pre-trained models are fed the raw input video and the natural language inquiry. In order to ensure that both modalities are represented in a rich and significant way, these pre-trained encoders extract basic visual and textual elements. The Feature Representation Module receives the extracted features and uses self-attention processes to represent the temporal linkages in the video and the contextual dependencies in the inquiry.

This stage aids the model in comprehending the relationships between various video frames and how the question's words add to its meaning.

4. FEATURE EXTRACTION

Two distinct pre-trained models are applied to both the question text and the raw video:

- Visual characteristics are extracted by a video backbone. either clip by clip or frame by frame.
- The question is transformed into a series of embeddings via a language model.

The first-level characteristics of the two modalities vision and language are these outputs.

5. FEATURE REPRESENTATION MODULE

Following extraction, the video and question features are processed separately by the model.

- The system learns relationships within each modality through self-attention.
- The relationship between frames over time in the video.
- Regarding the query: the relationships between various words.

Before merging the two modalities, this stage fortifies their internal organization and meaning.

6. QUESTION LOCALIZATION MODULE

This is the core innovation of LocAns. Using cross-modal attention and bi-linear fusion, the system merges the enriched visual features with the question representation. The combined features allow the model to:

- Understand how the question interacts with each segment of the video.
- Predict which temporal region of the video the question refers to.

The module generates multiple proposals (temporal segments) and scores them to identify the most likely segment that contains the answer.

7. ANSWER PREDICTION MODULE

The model clips the fused features to retain only that segment after determining which part is pertinent. The system forecasts the response from the candidate set using this modified input. This guarantees that irrelevant frames won't affect the final prediction.

8. TRAINING STRATEGY

Real temporal annotations—the precise timestamps that show where the solution is in the video—are unavailable for the majority of datasets, which presents a training issue. As a result, completely supervised training of the question localization module is not possible. The authors present pseudo-temporal labels, which are created automatically during training, to get around this restriction. The localization module is roughly supervised by these pseudo-labels. Further-more, a decoupled alternative training approach is used to train the model, where: The answer prediction module and the localization module are updated independently.

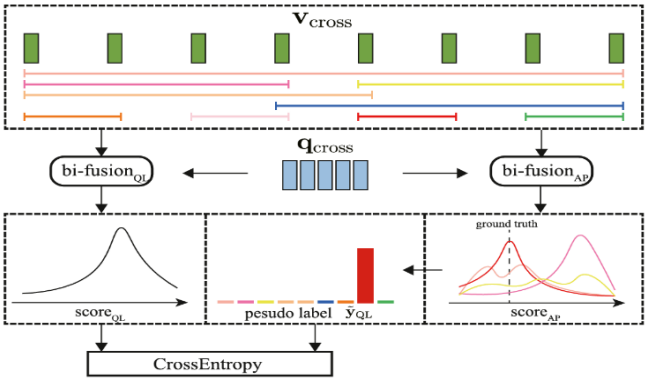

Fig. 2. Overview of Pseudo Label Generation and Question Localization Loss (LQL).

Fig. 2 provides an illustration of the concept. We assess each of the N temporal proposals produced for the input video during training. The answer prediction module processes each suggestion after it has been paired with the question. The most pertinent option for that question is presumed to be the one that yields the highest prediction score for the right response. After that, this suggestion is chosen as the pseudo-label, which serves as the localization module's supervisory signal.

To put it simply, the localization module is trained using the segment that the answer prediction module helps us determine is most likely to contain the answer. This pseudo-labeling technique enables the framework to directly assist the learning of question localization, greatly enhancing the model's capacity to identify the proper temporal location in lengthy movies, even though it is not a perfect replacement for actual human-annotated labels.

9. RESULT AND DISCUSSION

We carried out extensive experiments on three popular and recently generated datasets—NExT-QA, ActivityNet-QA, and AGQA—to assess how well our pro-posed LocAns architecture handles long-term Video Question Answering (Vide-oQA). Because the videos in these datasets are far longer and more semantically complex than those in previous VideoQA standards, they are particularly appropriate for our investigation.

10. DATASET OVERVIEW AND RATIONALE

Real-world films gathered from the YFCC-100M dataset and manually annotated for question-answering tasks are included in NExT-QA. It consists of:

5,440 videos, each lasting 44 seconds on average.

52,044 pairs of questions and answers.

Each question has five possible answers, with only one right response.

This configuration is perfect for evaluating the model's capacity to select the most contextually appropriate option since it permits controlled, multiple-choice reasoning.

The ActivityNet video dataset's descriptions are automatically re-annotated to create ActivityNet-QA. Important figures consist of:

5,800 videos. An average video lasts 180 seconds, or three minutes.

58,000 QA pairings divided into 32k train, 18k val, and 8k test.

There are no predetermined incorrect answers in this dataset; just the right response is given. We create an answer vocabulary using the top 1,000 most common responses in the training set in order to transform this into a classification task.

Additionally, 4,883 QA pairings that don't fit into this vocabulary are eliminated in this step. This dataset evaluates the model's performance on lengthier films with significant background scenes, which makes localization even more important.

The purpose of AGQA is to evaluate deeper reasoning skills pertaining to com-positional spatiotemporal relationships. Features of AGQAv2

9,700 videos

2.27 million pairs of questions and answers 30 seconds on average

Videos need a much greater level of reasoning sophistication even though they are shorter than ActivityNet. Because of this, AGQA is an effective dataset for assessing high-level relational and temporal reasoning.

11. IMPACT OF ATTENTION LAYERS ON MODEL PERFORMANCE

Analyzing how the quantity of self-attention and cross-attention layers impacts the final VideoQA performance is a crucial component of our research. Figure 3 summarizes this, and the following are our findings:

Trends in Performance

Early on, increasing the number of attention layers greatly improves performance.

Performance starts to plateau after it reaches a particular depth.

After this, adding more layers marginally reduces accuracy, indicating overfitting or an overly complex model.

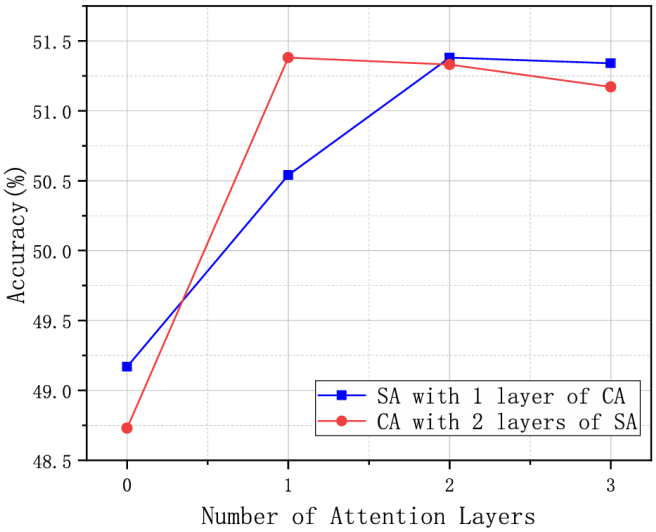

Fig. 3. Performance comparison of LocAns on the NExT-QA dataset when varying the number of self-attention (SA) and cross-attention (CA) layers.

Figure 3 shows how varying the number of self-attention (SA) and cross-attention (CA) layers affects the LocAns model's performance on the NExT-QA dataset.

This analysis aims to determine the level of attention depth required to capture visual-temporal information as well as the interaction between the question and the video.

Initially, performance is greatly enhanced by adding more self-attention layers. This is due to the fact that self-attention enables the model to identify long-range relationships between frames and comprehend the temporal flow of the movie. Adding more self-attention layers, however, eventually results in little to no improvement and even a slight decline in performance. This implies that overfitting, in which the model becomes overly specialized to training data and loses its capacity for generalization, is caused by deeper attention stacking.

Cross-attention layers show a similar trend. Because it directly matches the inquiry with the pertinent visual content in the film, cross-attention is essential. The accuracy of the model drastically decreases when the number of cross-attention layers is eliminated. The model's inability to successfully connect queries to visual cues in the absence of cross-attention is evident from this performance difference.

The model achieves the optimal balance between representational power and generalization when it employs two layers of self-attention and one layer of cross-attention.

Overall, the picture emphasizes the significance of both attention mechanisms and demonstrates that careful balancing of the architecture is necessary to attain the best outcomes.

12. CONCLUSION AND FUTURE WORK

In this research paper, we presented "locate before answering," a novel and successful paradigm for lengthy Video Question Answering (Video QA). Our paradigm first determines where in the film the answer is likely to be located and then concentrates the reasoning process solely on that localized section, in contrast to existing approaches that try to evaluate a full long video—which frequently contains a mixture of relevant and irrelevant sequences. This approach directly addresses the challenges presented by lengthy movies, where answer prediction is difficult due to intricate visual transitions, numerous activities, and scene changes. Through a novel pseudo-labeling technique, the two modules are concurrently tuned, enabling efficient learning even in the absence of explicit temporal ground-truth labels. Extensive tests on three contemporary long-term VideoQA benchmarks—NExT-QA, ActivityNet-QA, and AGQA—show that LocAns regularly performs better than current techniques. The advantages of explicitly localizing the question-relevant segment before responding are confirmed by both quantitative measurements and qualitative representations.

Our method increases overall accuracy, decreases noise from unrelated visual content, and improves the model's interpretability. Future research should continue to focus on closing this gap by creating or utilizing temporally annotated benchmark datasets. Furthermore, the "localization-first" approach put forward here has potential uses outside of Video QA and could spur improvements in longer-term video comprehension tasks including procedural event analysis, video grounding, and summarization.

References

1. Y.Wang, J.Deng, W.Zhou,and H.Li,“Weakly supervised temporal adja cent network for language grounding,” IEEE Trans. Multimedia, vol. 24, pp. 3276–3286, 2022.
2. L. H. Dang, T. M. Le, V. Le, and T. Tran, “Hierarchical object-oriented spatio-temporal reasoning for video question answering,” in Proc. Int. Joint Conf. Artif. Intell., 2021, pp. 636–642
3. Sharma, A., Gupta, A.K., Yadav, D., Barua, T. (2023). Optimizing Water Quality Parameters Using Machine Learning Algorithms. In: Marriwala, N., Tripathi, C., Jain, S., Kumar, D. (eds) Mobile Radio Communications and 5G Networks. Lecture Notes in Networks and Systems, vol 588. Springer, Singapore. https://doi.org/10.1007/978-981-19-7982-8_53.
4. W. Zhang et al., “Frame augmented alternating attention network for video question answering,” IEEE Trans. Multimedia, vol. 22, no. 4, pp. 1032–1041, Apr. 2020.
5. Gowda, V. Dankan, Sharma, Avinash, Kumaraswamy, S., Sarma, Parismita , Hussain, Naziya, Dixit, Santosh Kumar & Gupta, Anand Kumar(2023) A novel approach of unsupervised feature selection using iterative shrinking and expansion algorithm, Journal of Interdisciplinary Mathematics, 26:3, 519-530, DOI: 10.47974/JIM-1678
6. K. Khurana and U. Deshpande, “Video question-answering techniques, benchmark datasets and evaluation metrics leveraging video captioning: A comprehensive survey,” IEEE Access, vol. 9, pp. 43799–43823, 2021.
7. J. Wang, B.-K. Bao, and C. Xu, “DualVGR: A dual-visual graph reason ing unit for video question answering,” IEEE Trans. Multimedia, vol. 24, pp. 3369–3380, 2022.
8. J.Lei,L.Yu,M.Bansal,andT.L.Berg,“TVQA: Localized,compositional video question answering,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2018, pp. 1369–1379.
9. J. Xiao, X. Shang, A. Yao, and T.-S. Chua, “Next-QA: Next phase of question-answering to explaining temporal actions,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 9777–9786.
10. Z. Yu et al., “ActivityNet-QA: A dataset for understanding complex web videos via question answering,” in Proc. AAAI Conf. Artif. Intell., 2019, pp. 9127–9134.
11. L. Anne Hendricks et al., “Localizing moments in video with natural language,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 5803–5812.
12. J. Gao, C. Sun, Z. Yang, and R. Nevatia, “Tall: Temporal activity local ization via language query,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 5267–5275.
13. M.Grunde-McLaughlin, R.Krishna, and M.Agrawala, “AGQA:A bench mark for compositional spatio-temporal reasoning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 11287–11297.
14. Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, “Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 12, pp. 5947–5959, Dec. 2018.
15. Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, “Deep modular co-attention net works for visual question answering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 6281–6290.

16. A. K. Gupta, A. Sharma, A. Srinivasulu, T. Barua, S. Rajeyyagari and M. Subramanyam, "Early Prediction of Breast Cancer through Deep RNN Approach," 2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT), Pune, India, 2022, pp. 1-4, doi: 10.1109/TQCEBT54229.2022.10041634.
17. Gupta, Anand Kumar, Srinivasulu, Asadi, Oyerinde, Olutayo Oyeyemi, Pau, Giovanni, Ravikumar, C. V., COVID-19 Data Analytics Using Extended Convolutional Technique, Interdisciplinary Perspectives on Infectious Diseases, 2022, 4578838, 10 pages, 2022. https://doi.org/10.1155/2022/4578838.
18. J. Pennington, R. Socher, and C. D. Manning, "GLOVE: Global vectors for word representation," in Proc. Conf. Empirical Methods Natural Lang. Process., 2014, pp. 1532–1543.
19. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2019, pp. 4171–4186.
20. Shankar R., Kumar I., Mishra R.K. (2019). Outage probability analysis of MIMO-OSTBC relaying network over Nakagami-m fading channel conditions, Traitement du Signal, Vol. 36, No. 1, pp. 59-64. https://doi.org/10.18280/ts.360108
21. Kumar, I., Mishra, M.K., Mishra, R.K. (2021). Performance analysis of NOMA downlink for next- generation 5G network with statistical channel state information. Ingénierie des Systèmes d'Information, Vol. 26, No. 4, pp. 417-423. https://doi.org/10.18280/isi.260410
22. Shankar, R., Kumar, I., Mishra, R.K. (2019). Pairwise error probability analysis of dual hop relaying network over time selective Nakagami-m fading channel with imperfect CSI and node mobility. Traitement du Signal, Vol. 36, No. 3, pp. 281-295. https://doi.org/10.18280/ts.360312
23. Kumar I, Kumar A, Kumar Mishra R. Performance analysis of cooperative NOMA system for defense application with relay selection in a hostile environment. The Journal of Defense Modeling and Simulation. 2022;0(0). doi:10.1177/15485129221079721.
24. Ashish, I. Kumar and R. K. Mishra, "Performance Analysis For Wireless Non-Orthogonal Multiple Access Downlink Systems," 2020 International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET), Patna, India, 2020, pp. 1-6, doi: 10.1109/ICEFEET49149.2020.9186987.
25. Kumar, I., Mishra, R.K. (2021). An investigation of spectral efficiency in linear MRC and MMSE detectors with perfect and imperfect CSI for massive MIMO systems. Traitement du Signal, Vol. 38, No. 2, pp. 495-501. https://doi.org/10.18280/ts.380229.
26. Anand Kumar Gupta, Asadi Srinivasulu, Kamal Kant Hiran, Tarkeswar Barua, Goddindla Sreenivasulu, Sivaram Rajeyyagari and Madhusudhana Subramanyam. Early prediction and analysis of mammary glands cancer through deep learning approaches. World Journal of Advanced Engineering Technology and Sciences, 2022, 06(01), 018–024. Article DOI: https://doi.org/10.30574/wjaets.2022.6.1.0056
27. Kumar, I., Mishra, R.K. (2020). An efficient ICI mitigation technique for MIMO-OFDM system in time-varying channels. Mathematical Modelling of Engineering Problems, Vol. 7, No. 1, pp. 79-86. https://doi.org/10.18280/mmep.070110
28. Patil, J., Prajapati, K., Patel, D., Chauhan, R., Patel, M. (2026). A Review of Transforming AI for Depression Detection: Transformer Model Dominance, Multimodal Approaches, and Future Pathways. In: Bansal, J.C., Borah, S., Hussain, S., Salhi, S. (eds) Computing and Machine Learning. CML 2025. Lecture Notes in Networks and Systems, vol 1612. Springer, Singapore. https://doi.org/10.1007/978-981-95-2872-1_7
29. Kumar I., Sachan V., Shankar R., Mishra R.K. (2018). An investigation of wireless S-DF hybrid satellite terrestrial relaying network over time selective fading channel, Traitement du Signal, Vol. 35, No. 2, pp. 103-120. https://doi.org/10.3166/TS.35.103-120
30. Sachan, V., Kumar, I., Shankar, R., Mishra, R.K. (2018). Analysis of transmit antenna selection based selective decode forward cooperative communication protocol. Traitement du Signal, Vol. 35, No. 1, pp. 47-60. https://doi.org/10.3166/TS.35.47-60
31. Indrajeet Kumar, Vikash Sachan, Ravi Shankar, Ritesh Kumar Mishra, Performance Analysis of Multi-User Massive MIMO Systems with Perfect and Imperfect CSI, Procedia Computer Science, Volume 167, 2020, Pages 1452-1461, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.356.
32. Vikash Sachan, Indrajeet Kumar, Lokesh Bhardwaj, Ritesh Kumar Mishra, Pairwise Error Probability Analysis of SM-MIMO system employing k – μ Fading Channel, Procedia Computer Science, Volume 167, 2020, Pages 2516-2523, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.304.
33. Srinivasulu, A., Gupta, A. K., Kolambakar, S. B., Subramanyam, M., Rajeyyagari, S. R., Barua, T., & Pushpa, A. (2022, October). Prostate Cancer Data Analytics Using Hybrid ECNN and ERNN Techniques. In International Conference on Business Data Analytics (pp. 36-52). Cham: Springer Nature Switzerland.
34. R. Shankar, I. Kumar, A. Kumari, K. N. Pandey and R. K. Mishra, "Pairwise error probability analysis and optimal power allocation for selective decode-forward protocol over Nakagami-m fading channels," 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), Chennai, India, 2017, pp. 1-6, doi: 10.1109/ICAMMAET.2017.8186700.
35. Gupta, Y., Verma, R., Sharma, S.S.P.M.B., Kumar, I. (2024). An IoT Application Based Decentralized Electronic Voting System Using Blockchain. In: Pareek, P., Gupta, N., Reis, M.J.C.S. (eds) Cognitive Computing and Cyber Physical Systems. IC4S 2023. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 537. Springer, Cham. https://doi.org/10.1007/978-3-031-48891-7_12

36. Mishra, N., Raghuwanshi, R., Maurya, N.K., Kumar, I. (2024). Efficient Fuel Delivery at Your Fingertips: Developing a Seamless On-Demand Fuel Delivery App with Flutter. In: Pareek, P., Gupta, N., Reis, M.J.C.S. (eds) Cognitive Computing and Cyber Physical Systems. IC4S 2023. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 537. Springer, Cham. https://doi.org/10.1007/978-3-031-48891-7_11
37. Trivedi, D., Saxena, M., Sharma B, S.S.P.M., Kumar, I. (2024). Harmonizing Insights: Python-Based Data Analysis of Spotify's Musical Tapestry. In: Pareek, P., Gupta, N., Reis, M.J.C.S. (eds) Cognitive Computing and Cyber Physical Systems. IC4S 2023. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 536. Springer, Cham. https://doi.org/10.1007/978-3-031-48888-7_3
38. A. Sharma, A. Srinivasulu, T. Barua and A. K. Gupta, "Deep Learning based Detection and Prediction of Omicron Diagnosis on Collected Symptoms," 2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2022, pp. 1258-1262, doi: 10.1109/ICCES54183.2022.9835882.
39. D. Xu et al., “Video question answering via gradually refined attention over appearance and motion,” in Proc. 25th ACM Int. Conf. Multimedia, 2017, pp. 1645–1653.