ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

VidSum: Deep Learning Video Summarization

[1]Dr. R Rambabu, [2]Yannam Lakshman Manidhar,[3]Thiramdasu Subrahmanyam, [4]R Siva Durga Sai Surendra,

[1] Professor, Department of CSE, Rajamahendri Institute of Engineering & Technology, Bhoopalapatnam, Near Pidimgoyyi, Rajahmundry, E. G. Dist. A.P 533107.

[2,3,4]Student, Department of CSE, Rajamahendri Institute of Engineering & Technology, Bhoopalapatnam, Near Pidimgoyyi, Rajahmundry, E. G. Dist. A.P 533107.

Abstract—

In recent years, the work on video-text retrieval has been well-developed due to the emergence of large-scale pre-training methods. However, these works focus solely on inter-modal interactions and contrasts, neglecting the contrasts of multi-grained features within modalities, which makes the similarity measurement less accurate. Worse still, many of these works only contrast features of the same grain size, ignoring the important information implied by features of different grain sizes. For these reasons, we propose a joint modal similarity contrastive learning model, named JM-CLIP. Firstly, the method employs a multidimensional contrastive strategy of two modal features, including inter-modal and intra-modal multi-grained feature contrasts. Secondly, to fuse the various contrastive similarities well, we also design a joint modal attention module to fuse the various similarities into a final joint multigranularity similarity score. The significant performance improvement achieved on three popular video-text retrieval datasets demonstrates the effectiveness and superiority of our proposed method.

INTRODUCTION

There is a lot of video data that we engage with and consume these days because of internet video services like Netflix and YouTube and social media. Hundreds of hours of video are posted to the internet every single minute. Beyond that, video data has multiplied in the last decade and is still growing at an incredible pace, all because cameras are everywhere, from our cellphones to inexpensive CCTV security cameras. Consequently, it is now more important than ever to develop effective video summarizing systems that may extract the most important parts of a video while also reducing the workload of human viewers. The ability to quickly access concise summaries of lengthy videos is a major benefit of video summarizing. Reducing the original video to a few key frames can make videos more educational, impressive, and engaging. There are several practical uses for video summarization in the real world. In this regard, video surveillance is an extremely relevant case. Humans aren't particularly efficient when it comes to watching video surveillance material that's been recorded for 24 or even 72 hours. Nevertheless, with Video Summarization, investigators just need to view the synopsis video; instead of watching hours of surveillance footage, they won't miss a thing—the model will identify crucial moments in hours of footage, such as that of a parking lot, and extract the crucial scenes to

RELATED WORK

Video summary has been approached from several angles. For video summarizing, Kanafani et al. [2] used Unsupervised Learning [16]. Similarly, Shemer et al. [13] and Jadon et al. [12] have used Unsupervised Learning to extract crucial moments from videos. Also, a lot of studies on video summarization have made use of supervised learning, including methods like recurrent neural networks [18], convolutional neural networks [17], etc. [20]. Video summaries have been created using supervised learning by Zhu et al. [4] and Elfeki et al. [15]. Researchers Huang et al. [14], Rochan et al. [10] and Fajtl et al. [11] have used CNNs, whereas Vasudevan et al. [9] have utilized both CNNs and recurrent neural networks. Video summarizing has also made use of some quite novel methods. Graphs, in which nodes represent scenes in a movie, have been used by Papalampidi et al. [7]. For the purpose of summarization, Hohman et al. [8] have collected text and colors from footage of popular TV episodes. When training their networks to generate video summaries, Apostolidis et al. [3] and Fu et al. [1] made use of General Adversarial Networks [22]. On the other hand, some academics have turned to Reinforcement Learning [21] as a solution to the video summarization issue. Zhou et al. [6] and Lan et al. [5] trained their deep learning models using

Reinforcement Learning. Video summarization has been trialed with every conceivable learning technique, including supervised, unsupervised, and reinforcement learning. However, supervised learning approaches are the most successful. Supervised learning was chosen over other approaches since the model performs much better in this particular circumstance when it comes to learning. Among the many methods tried to address the issue of video summarization, supervised learning has produced some of the most promising outcomes so far. Additionally to Convolutional Neural Networks, we have included Long Short Term Memory (LSTM) Networks [19] into our model. We opted for LSTM networks instead of other possibilities because: A. RNs have the issue of vanishing gradients, which prevents them from learning from extended data sequences. Because of their close relationship to the forget gate's activation function, LSTM are able to resolve the Vanishing Gradient Problem. B. On huge datasets, LSTM Networks outperform Gated Recurrent Units (GRUs) [23]. When it comes to datasets with massive quantities of data, they nail it. C. Convolutional Neural Networks (CNN) are limited to finding spatial characteristics and patterns in picture and audio data; LSTM Networks, by their very nature, can analyze sequential data and provide predictions based on it.

PROPOSED APPROACH

First, we recommend breaking a video into its individual frames. The Convolutional Neural Network takes in the video frames and uses its Convolutional Layers to pull out spatial information. In order to assess the time-based relevance of these frames, the LSTM network is given these extracted characteristics. The next step is to determine whether each picture will be included in the final summary by calculating their Temporal Intersection over Union with the ground truth. whether so, they are classed as yes. The last stage is to construct the summary movie by compiling all the photos that were classed as yes. There are five stages to our suggested solution: A. De-duplicating Video Frames: Multiple video sequences comprise a single video, and each video sequence in turn contains multiple pictures. Therefore, we start by segmenting the footage. The next step is to pull stills from the movie. The result is a visual depiction of a video consisting of all its individual frames. Here, we use the Python package OpenCV. The majority of CCTV footage is 24 frames per second (fps), which translates to 24 pictures per second. Consequently, the model is fed 24 pictures per second culled from the video. Other frame rates, such as 30 and 60 frames per second, are also supported by our model. B. Extracting characteristics from the Extracted Frames: With these frames as a starting point, our model can learn the characteristics. We do this by extracting spatial characteristics from the photos using convolutional neural networks. Following the aforementioned process of video frame extraction, our model is now ready to begin learning. To prevent overfitting, we have further included dropout layers in the model. The characteristics that our model has retrieved from the picture frames are given to us via the output of these layers.

C. Plans for the future: Identifying the most crucial frames to include in the summary is the next step. Here, we use Long Short Term Memory (LSTM) cells. We rely on LSTM since it can also retrieve data from earlier LSTM cells. The time-based interest video segments are crucial to the success of any film or video. It is necessary to know what scene came before a scene in order to establish the scene's importance. Thanks to LSTM, our model has "memory" and can recall prior scenes, allowing it to more accurately identify which parts of the video are most relevant to the summary. Next, the model uses LSTM to suggest, in a time-coherent fashion, which frames are interesting and significant. The last step is to determine whether a frame will be included in the final summary by using the classification based on the intersection of time and union. We consider a proposal to be positive and assign it a score of 1 if its Temporal Intersection over Union (tIOU) is higher than 0.6. If this value is more than the Ground Truth summary frames provided in the dataset, we incorporate it in the final summary. Every frame is deemed negative and removed from the final summary if its Temporal Intersection over Union value is below 0.6. Therefore, the Temporal Intersection Over-Union Threshold remained at 60% during model training. Section E. Frame Combination: To create the summary video, you must collect all the frames that are good and have been awarded a score of 1. The output is the final video summary, which is achieved by combining these frames using the OpenCV library. Figure 1 shows the flow diagram for our proposed method, VidSum - video summarization.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

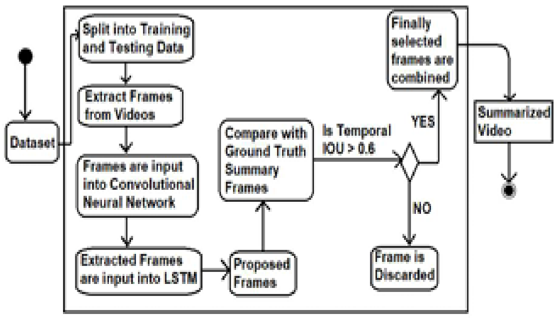


Fig. 1. Proposed Approach Flowchart of VidSum - Video Summarization using Deep Learning

IMPLEMENTATION

E. Inputting these frames into LSTM: The procedures we used to execute our approach are as follows: A. Gathering the Datasets: We need training data in order to train our model. To do this, we need video files in addition to ground truth summaries that have been human-created. In order for the model to learn and recognize the most relevant frames in a video, as well as to verify that the recommended frames intersect with the ground truth frames in terms of temporal intersection over union, we need this. As a result, we gathered the well-known TVSum [24] and SumMe [25] datasets, which are used for video summarizing and feature human summaries for each video across genres. To validate our model, we also require testing data to evaluate its performance in terms of summary frame proposal and temporally coherent summary creation. To accommodate both training and testing needs, we partitioned the TVSum and SumMe datasets. We organized all the video clips and their human-created ground truth user summaries after checking for corrupted files. Doing so is critical for the model's learning and training on the input videos. We developed a Python program to extract picture frames from the video as part of the preprocessing step for the input video files (B). For this, we relied on the OpenCV Python package. The goal is for the Convolutional Neural Network to be able to use these extracted frames as training data. C. Feature Detection and Extraction: We developed a unique ML model using Maxpool2D layers interspersed with Convolutional layers. In order to improve its prediction of which frame to use for the summary video, the Convolutional Neural Network trains on spatial information extracted from the photos. We train our CNN on the input video frames and then extract features from them, as shown in Fig. 2.

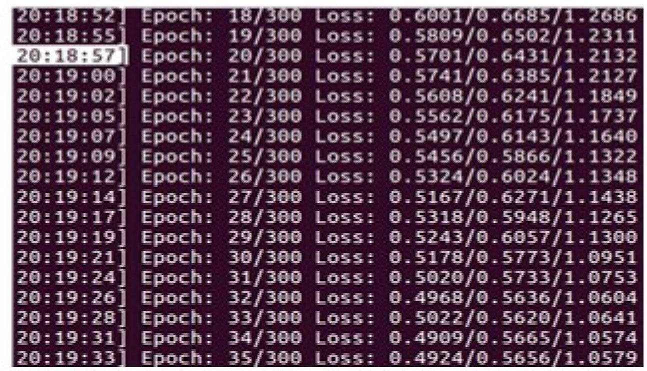
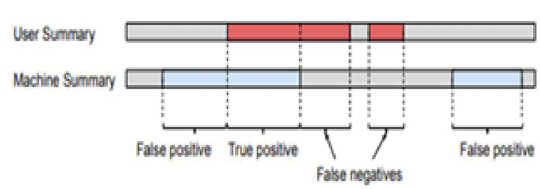


Fig. 2. Convolutional Neural Network training and extracting features from the frames

D. Minimizing Overfitting: Since our model was overfitting, we interspersed the convolutional layers with dropout layers that had a probability of 0.5. By doing so, overfitting is lessened. Our model has retrieved spatial information from the picture frames, and these layers' output offers us those features.

E. Inputting these frames into Long Short-Term Memory (LSTM): Now that we know how to extract spatial features using convolutional layers, we can use LSTM to determine which frames are coherent in terms of time. The "Memory" feature of LSTM allows it to recall the scene that came before the current one, allowing it to assess the significance of the current frame. In order to create the summary movie, the LSTM layers provide output frames that they believe should be included. We refer to them as frameworks for proposals based on temporal interests. F. Matrix comparison with ground truth summary video frames: Figure 3 shows the results of a temporal intersection over union calculation performed on the suggested frames and the ground truth summary video frames from the dataset.



INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

Fig. 3. Comparing Proposed Frames with Ground Truth Summary Frames

G. Arranging these frames according to their Temporal IOU classification: We mark a frame as positive and include it in the final summary movie if its Temporal Intersection over Union is greater than 0.6. For a given frame, if the TIMOV of the merge is

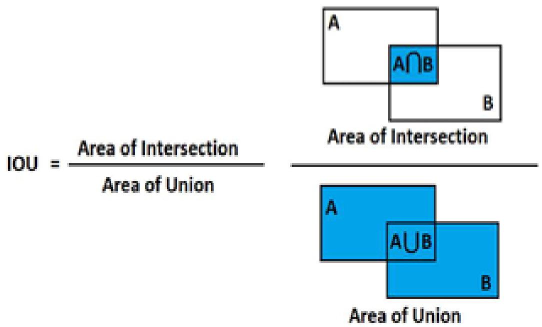


Fig. 4. Calculating Temporal Intersection over Union

To create the summary video, we include all frames with a Temporal Intersection Over Union (tIOU) greater than 0.6 and exclude any frames with a tIOU less than 0.6. The next step is to merge the frames. To create the summary video, all of the chosen frames are composited in the proper sequence. The OpenCV python package is what we're using here. Getting the model ready: We spent 30 hours training the model with 10,000 iterations. In order to improve the model's accuracy, we trained it again after making many adjustments. To prevent overfitting, we included several dropout layers. The goal was to improve the model's ability to understand the crucial parts of a video and make it more generalizable. Our model is trained on the SumMe dataset (Fig. 5) and the TVSum dataset (Fig. 6).

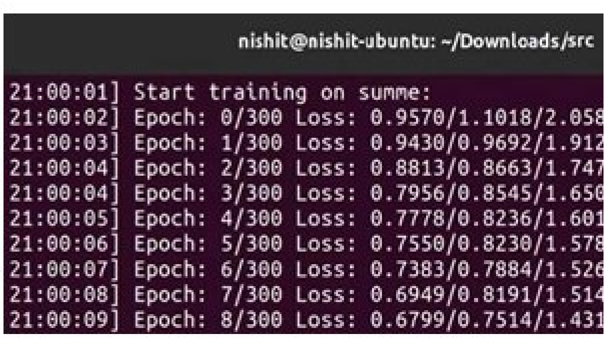


fig. 5. Training our model on SumMe Dataset

Fig. 6. Training our model on TVSum Dataset

TESTING AND OBSERVATIONS

As a first step in evaluating the model, we ran it against the testing datasets that were initially created when we divided the TVSum and SumMe datasets for training and testing. In order to improve the model's accuracy, we adjusted its hyperparameters and made other fine-tuning adjustments. We tweaked the model several times and fine-tuned it more to boost the

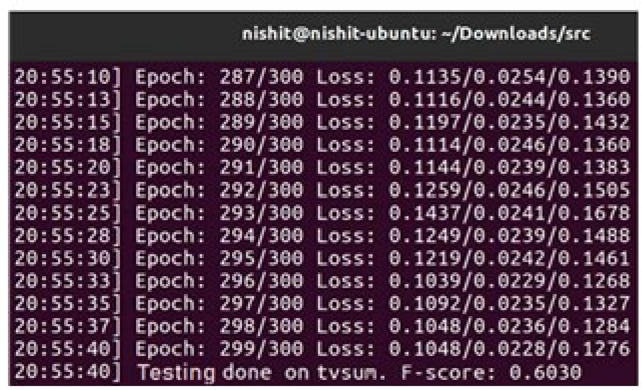


Fig. 7. Testing our model on TVSum Dataset

preciseness of the model. In the end, we were able to get testing F-Scores of 60.3% on the TVSum Dataset and 48.9% on the SumMe Dataset, as seen in Figures 7 and 8, respectively.
Part B: Remarks: Different Time-Based IOU Cutoffs: In this situation, we tried adjusting the Temporal IOU Threshold from 0.6 to 0.7 and 0.4, however the resulting summary video lacked temporal coherence and yielded poor results. We discovered that setting

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

the Temporal IOU Threshold to 0.6 yielded the most effective summary films. Furthermore, as seen in Figure 9, the maximum F-Score was obtained with a Temporal Intersection Over Union criterion of 0.6. Consequently, the optimal value for the Temporal IOU Threshold was 0.6.

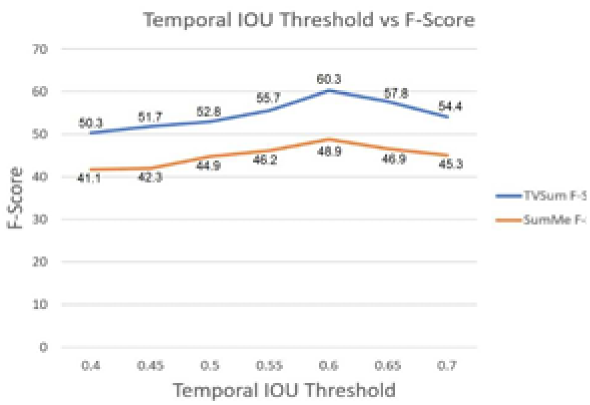


Fig. 9. Temporal IOU Threshold vs F-Score

RESULTS

Finally, we were able to get testing F Scores of 60.3% on the TVSum Dataset and 48.9% on the SumMe Dataset after fine-tuning the model and training it for 10,000 iterations over 30 hours. In terms of Video Summarization, they represent an improvement over earlier methods. Comparisons with prior research on video summarization using the TVSum and SumMe datasets are shown in Table I, along with our F-Score.

TABLE I. OUR FINAL F-SCORE COMPARED TO PREVIOUS WORKS

Model	TVSum F-Score	SumMe F-Scor
FCSN (Rochan et al.) [10]	52.7	41.5
VASNet (J. Fajtl et al.) [11]	59.8	47.7
DR-DSN (Zhou et al.) [6]	57.6	41.4
Ours	60.3	48.9

Final Testing F-Scores achieved by our model on the TVSum and SumMe Dataset are shown in Fig. 10 and Fig. 11 respectively.

Fig. 10. Final Testing F-Scores of our model on TVSum Dataset

Fig. 11. Final Testing F-Scores of our model on SumMe

Figure 12 shows the Dataset F-Scores obtained by our model on the SumMe Dataset, while Figure 13 shows the F-Scores attained on the TVSum Dataset, in comparison to existing Video Summarization approaches. Figure 14 displays still images from a video summary that our algorithm generated from security camera data.

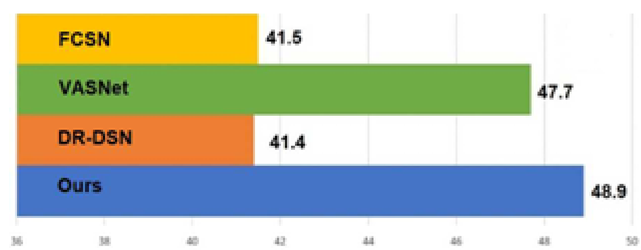
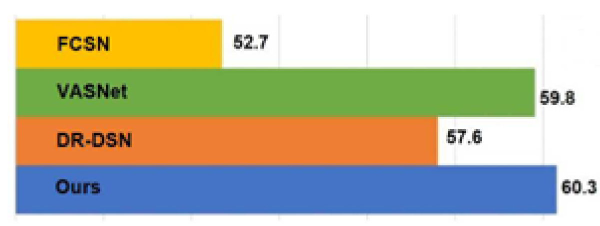


Fig. 12. Our Final F-Score as compared to other video summarization techniques on SumMe Dataset



INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

Fig. 13. Our Final F-Score as compared to other video summarization techniques on TVSum Dataset

CONCLUSION

this document so lays out our system for summarizing videos. Our approach views video summarization as a problem of temporal interest detection, in contrast to existing algorithms that learn the relevance score of each frame in the movie independently. This allows it to learn the temporal coherence between video frames, which improves its performance compared to existing supervised models when it comes to video summarization. Ultimately, our model improved to the point where it could properly extract the most crucial parts of a video clip and put them in the summary video while ignoring the rest. Since our model could effectively generate a high-quality summary video including all the crucial elements of a video clip, we were able to enhance accuracy for the video summarizing issue. Our goal moving forward is to make our unified structure more interest-based coherent. If we want our model to work better, we might also try to make the transitions between scenes in a film more seamless in time.

Fig. 14. Screenshots from summary video created by our model on CCTV surveillance footage

REFERENCES

[1]. T. -J. Fu, S. -H. Tai and H. -T. Chen, "Attentive and Adversarial Learning for Video Summarization," 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2019, pp. 1579-1587, doi: 10.1109/WACV.2019.00173.

[2]. Kanafani, Hussain, et al. Unsupervised Video Summarization via Multi-Source Features. arXiv, 26 May 2021. arXiv.org, https://doi.org/10.48550/arXiv.2105.12532.

[3]. E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris and I. Patras, "AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 8, pp. 3278-3292, Aug. 2021, doi: 10.1109/TCSVT.2020.3037883.

[4]. W. Zhu, J. Lu, J. Li, and J. Zhou. 2021. DSNet: A Flexible Detect-to Summarize Network for Video Summarization. Trans. Img. Proc. 30 (2021), 948–962. https://doi.org/10.1109/TIP.2020.3039886

[5]. S. Lan, R. Panda, Q. Zhu, A.K. Roy-Chowdhury. (2018). “FFNet: Video Fast-Forwarding via Reinforcement Learning”. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6771-6780. unsupervised

[6]. K. Zhou, Y. Qiao, and T. Xiang. 2018. “Deep reinforcement learning for video summarization with diversity representativeness reward”. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18). AAAI Press, Article 929, 7582–7589.

[7]. P. Papalampidi, F. Keller, and M. Lapata, “Movie Summarization via Sparse Graph Construction”, AAAI 2021, 2021

[8]. Fred Hohman, Sandeep Soni, Ian Stewart, and John Stasko, 2017, A Viz of Ice and Fire: Exploring Entertainment Video Using Color and Dialogue, VIS4DH Workshop 2017

[9]. Arun Balajee Vasudevan, Michael Gygli, Anna Volokitin, and Luc Van Gool, 2017, Query-adaptive Video Summarization via Quality-aware Relevance Estimation, MM '17: Proceedings of the 25th ACM International Conference on Multimedia, https://doi.org/10.1145/3123266.312329720 17 pp. 582–590,

[10]. Mrigank Rochan, Linwei Ye and Yang Wang, 2018, Video Summarization

Using Fully Convolutional Sequence Networks, ECCV 2018, 2018

[11]. Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino, 2019, Summarizing Videos with Attention

[12]. Shruti Jadon, and Mahmood Jasim, 2020, Unsupervised video summarization framework using keyframe extraction and video skimming, 2020 IEEE 5th International Conference on Computing Communication and Automation 10.1109/ICCCA49541.2020.9250764 (ICCCA), 2020, doi:

[13]. Yair Shemer, Daniel Rotmanand, and Nahum Shimkin, 2019, ILS SUMM: Iterated local search for unsupervised video summarization, 2020 25th International Conference on Pattern Recognition (ICPR), 2020, doi: 10.1109/ICPR48806.2021.9412068

[14]. Jia-Hong Huang, and Marcel Worring, 2019, Query controllable video summarization, ICMR '20: Proceedings of the 2020 International Conference on Multimedia Retrieval, June 2020, pp. 242–250, doi: 10.1145/3372278.3390695

[15]. Mohamed Elfeki, Liqiang Wang, and Ali Borji, 2019, Multi stream dynamic video summarization, WACV 2022, pp. 185-195

[16]. Muhammad Usama, Junaid Qadir, Aunn Raza, and Hunain Arif, 2019, Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges, IEEE Access (Volume: 7), doi: 10.1109/ACCESS.2019.2916648

[17]. Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi, 2017, Understanding of a convolutional neural network, 2017 International Conference on Engineering and Technology (ICET), doi: 10.1109/ICEngTechnol.2017.8308186

[18]. Alex Sherstinsky, 2018, Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network, 2018

[19]. Greg Van Houdt, Carlos Mosquera, Gonzalo Nápoles, 2020, A Review on the Long Short-Term Memory Model, Springer Artificial Intelligence Review, Dec 2020, doi: 10.1007/s10462-020-09838-1

[20]. Vladimir Nasteski, 2017, An overview of the supervised machine learning methods, Dec 10.20544/HORIZONS.B.04.1.17.P05

[21]. Agrawal, K. K. ., P. . Sharma, G. . Kaur, S. . Keswani, R. . Rambabu, S. K. .Behra, K. . Tolani, and N. S. . Bhati. "Deep Learning-Enabled Image Segmentation for Precise Retinopathy Diagnosis". *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 12s, Jan. 2024, pp. 567-74, https://ijisae.org/index.php/IJISAE/article/view/4541.

[22]. Samota, H. ., Sharma, S. ., Khan, H. ., Malathy, M. ., Singh, G. ., Surjeet, S. and Rambabu, R. . (2024) "A Novel Approach to Predicting Personality Behaviour from Social Media Data Using Deep Learning", *International Journal of Intelligent Systems and Applications in Engineering*, 12(15s), pp. 539–547. Available at: https://ijisae.org/index.php/IJISAE/article/view/4788