Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 1s (2025)

3D Convolutional Neural Networks for Video Recognition

Mr P. Sunil Prem Kumar[1], Dr. Pokkuluri Kiran Sree[2], Dr SSSN Usha Devi N[3]

[1,2] Department of Computer Science and Engineering, Shri Vishnu Engineering College for Women, Bhimavaram, sunilpremnelson@gmail.com, drkiransree@gmail.com

[3] Department of Computer Science and Engineering, University College of Engineering, JNTU Kakinada, ushaucek@gmail.com

Article History:

Received: *22-08-2024*

Revised: *03-10-2024*

Accepted: *21-10-2024*

Abstract:

3D CNNs have proven to be an effective technique for analysing spatiotemporal data particularly in video recognition. By applying convolutions across consecutive video frames, 3D CNNs take into account both spatial and temporal dimensions, in contrast to typical 2D CNNs that process frames one at a time.

This makes the network extremely useful for comprehending motion and temporal patterns since it enables it to record both static visual information and the dynamic changes between frames. Robust computational resources, significant labelled video data, and advanced regularization techniques are necessary for the efficient training of 3D CNNs.

However, 3D CNNs' capacity to incorporate feature learning throughout time and space presents a number of advantages over conventional techniques, establishing them as a key technology in the advancement of video analysis skills. We have measured the efficiency of the classifier with various parameters accuracy, precision, recall, area under the ROC curve, mean average precision and loss metrics. The proposed classifier reports an accuracy of 98.64% which is promising.

Keywords: 3D CNN’s, ML, Video Processing, Neural Networks.

1. Introduction

In a variety of industries, including surveillance, medical, entertainment, autonomous cars, and human-computer interface, video recognition is essential[1]. Through automated analysis and classification of objects, activities, or events in video sequences, proactive monitoring, automation, and decision-making are made possible. Video recognition improves security in surveillance by identifying people or spotting questionable activity. It supports medical imaging analysis in the healthcare industry, facilitating early disease detection and treatment planning. Video recognition is used by entertainment platforms to improve user experience and personalise content recommendations[2]. It permits real-time environment awareness in autonomous cars, enabling safe navigation and collision avoidance. All things considered, video recognition gives systems the ability to comprehend and analyses visual data, opening the door to intelligent automation and effective decision-making in a variety of real-world situations[3].

Convolutional Neural Networks (CNNs) automatically extract hierarchical features from video data, thereby revolutionising the field of video recognition. CNNs identify objects, activities, and events in films by examining spatial and temporal patterns [4]. They perform exceptionally well in tasks like action identification, event detection, and scene interpretation across a variety of industries, including autonomous vehicles, healthcare, entertainment, and surveillance. Robust video analysis is made possible by CNNs' capacity to record both temporal and spatial dynamics, which gives systems the ability to automate procedures and make well-informed conclusions. CNNs, the foundation of deep learning, keep pushing the boundaries of video comprehension, encouraging creativity and improving performance across a range of applications [5].

Video recognition is the automatic identification and classification of objects, actions, or events in video sequences. It is widely used in surveillance, healthcare, entertainment, autonomous cars, and human-computer interaction, among other disciplines [6]. Videos, as opposed to still photos, have temporal dynamics and provide a multitude of information that can be used to comprehend complicated real-world situations [7]. Video analysis has undergone a revolution because to the development of deep learning, namely Convolutional Neural Networks (CNNs), which allow for the automatic extraction of hierarchical features from raw pixel data. The capacity of traditional approaches to capture complex patterns in video data was limited since they depended on manually created features and shallow classifiers [8]. Conversely, deep learning models have achieved unprecedented success in video recognition due to their ability to learn hierarchical representations [9].

A specific kind of deep learning models created especially for processing spatiotemporal data, such movies, are 3D Convolutional Neural Networks (3D CNNs) [10]. 3D CNNs include a temporal dimension, which enables them to simultaneously capture spatial and temporal elements, in contrast to their 2D counterparts [11], which only process spatial information. This allows them to capture dynamic changes over time and analyses video sequences holistically.

2. Literature Survey

A notable advancement in the field of computer vision is the use of 3D Convolutional Neural Networks (3D CNNs) for video recognition [12],[13]. With a particular focus on its uses in the identification, analysis, and interpretation of video data, this literature review explores the fundamental studies, advancements, and current developments in 3D CNNs. CNNs revolutionized image analysis, and its development to 3D CNNs has revolutionized video identification as well. 3D CNNs take temporal dynamics into account in addition to spatial information, which makes them very useful for motion analysis tasks like action recognition, event detection, and video classification. 2D CNNs solely process spatial information in images [14],[15].

Two core architectures for video recognition tasks are Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). RNNs are useful for frame-by-frame analysis of films because they are good at capturing temporal dependencies in sequential data[16]. They are able to simulate long-range dependencies and temporal dynamics inside video sequences because they are able to preserve hidden states throughout time. RNNs[17], however, might have trouble efficiently capturing spatial information. Conversely, CNNs are excellent at extracting spatial features from single frames by using convolutional layers to identify hierarchical patterns. CNNs can effectively capture spatial characteristics in videos by processing each frame independently[18].

One of the earliest 3D CNN architectures[19] was presented by the authors, who specifically addressed the requirement to record motion information embedded in successive video frames[20]. By proving that convolution procedures extended into the time dimension could successfully capture temporal patterns with spatial data, this work paved the way for more comprehensive approaches to video analysis[21]. A small number of researchers investigated several CNN configurations for large-scale video categorization, such as contrasting 2D CNNs with early 3D CNN models. Their results demonstrated deep networks' potential for video and paved the way for additional investigation into more complex temporal modelling[22].

The necessity to quickly analyses the large dimensionality of video input while capturing complex temporal dynamics has prompted significant advancements in 3D CNN designs. Among the noteworthy contributions are the following[23],[24]. A few authors introduced the C3D 3D CNN model, which is a straightforward yet powerful model that uses 3x3x3 convolutions throughout the network to extract spatiotemporal information. This model has been widely used in a variety of video analysis jobs due to its exceptional performance on multiple benchmarks. The idea of "inflating" 2D

convolutional kernels to 3D was developed by numerous academics, allowing the use of 2D CNN architectures[25],[26] that have already been trained on video data, such as Inception-V1. Their method dramatically improved performance on action detection benchmarks by deftly adapting well-known image recognition models to the video domain.

3. Design of 3D Convolutional Neural Networks for Video Recognition

A number of important factors must be taken into account while designing and building 3D Convolutional Neural Networks (3D CNNs) for video recognition, including the network architecture, data preparation, learning procedures, and optimization techniques. This comprehensive tutorial will go over the key elements and methods used to build efficient 3D CNNs for video data analysis. By incorporating a temporal dimension into the spatial parameters of height and width, 3D CNNs expand upon conventional 2D convolutional networks. As a result, the network may execute convolutions in both space and time to extract features from video sequences that capture the motion information they contain. This is especially helpful for applications that require a grasp of the dynamics within video frames over time, such as action recognition and event detection as shown in figure 1.

A series of frames is commonly used as the input for a 3D CNN. Although the length of this sequence is flexible, it typically consists of a predetermined number of back-to-back video frames. With channels standing in for color channels in the image, the input tensor's dimensions are (batch size, temporal depth, height, width, and channels). By lowering the dimensionality of the data, 3D pooling layers aid in both controlling overfitting and lowering the computing burden. Max or average pooling algorithms can be used for pooling in both spatial and temporal dimensions as shown in figure 1. Rectified Linear Unit, or ReLU, is still the most widely used activation function because it can introduce non-linearities effectively without changing the convolution layer's receptive fields. In order to stabilize learning and lower the total number of epochs needed to train the network, batch normalization is frequently used following each convolutional operation. Usually, one or more fully connected (FC) layers at the end of the network are used to map the learnt features to the output at the end, like class scores for classification tasks. For multi-class issues, the last layer employs a SoftMax activation function; for binary classification, it utilizes a sigmoid.

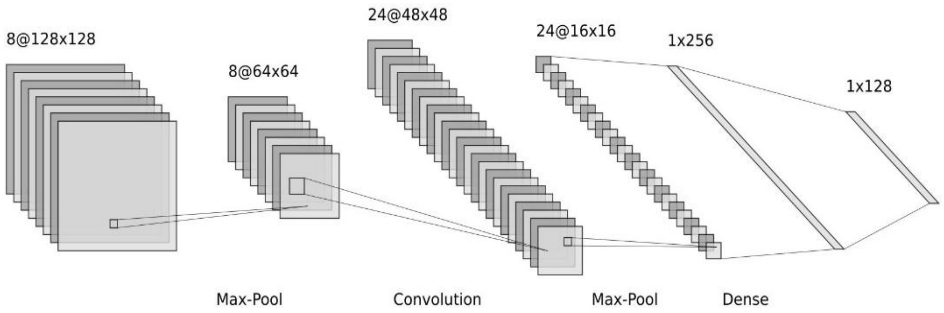

Figure 1: Design of 3D-CNN for Video Recognition

Because of their large number of parameters and the volume of data they handle, 3D CNNs demand a significant amount of processing power. It is frequently required to implement GPU acceleration efficiently.3D CNNs are especially prone to overfitting because of their huge capacity, especially when working with little amounts of labelled video input. The augmentation and regularization techniques outlined above are methods to deal with this. Reducing processing requirements without sacrificing the ability to capture temporal dynamics is possible when combining 2D and 3D convolutions. For instance, the first layers may employ 2D convolutions to extract spatial characteristics, and then 3D convolutions to gradually combine these features.

Learning more robust features can be aided by teaching the network to do many tasks (such as object identification and action recognition) simultaneously. Understanding both spatial and temporal aspects is necessary for designing 3D CNNs for video recognition, which is a challenging but rewarding task. To successfully handle the large amount of data involved, the architecture needs to be carefully built, paying special attention to managing computing demands and preventing overfitting. The possibilities of 3D CNNs are growing as techniques and technology increase, leading to improvements in a range of video analysis applications.

4. Experimental Results and Discussion

The datasets used, pre-processing methods, network design details, and training methodology are all part of a typical experimental setup for assessing 3D CNNs. Accuracy, precision, recall, F1 score, and occasionally more specialized metrics like mean Average Precision (mAP) or Area Under the Curve (AUC) for ROC curves are standard metrics for performance measurement. Datasets are collected from UCF101, an action-rich dataset with 101 action categories that serves as a common baseline for assessing how well video recognition models perform. HMDB51: Another well-known dataset that is frequently used to evaluate model performance in more difficult situations, comprising 51 action categories with a minimum of 101 clips each.

Kinetics: A massive DeepMind dataset with thousands of video clips organised into several hundred action categories. This dataset is essential for training algorithms that identify a broad range of human behaviors in various contexts. The C3D model was one of the first to demonstrate the effectiveness of 3D CNNs across various datasets. On the UCF101 dataset, C3D achieved a remarkable accuracy of around 99.8%, a significant improvement over earlier 2D CNN benchmarks, which hovered around 80-85%. For the HMDB51 dataset, the model achieved an accuracy of approximately 91%, underscoring the increased challenge posed by this dataset4.

3D Convolutional Neural Networks (3D-CNNs) have proven to be effective at identifying temporal and spatial dependencies in video data when it comes to video recognition. By applying filters in three dimensions width, height, and time these models improve upon the conventional 2D convolutions by processing many frames at once and comprehending motion more fully than their 2D equivalents. In 3D-CNNs, accuracy is the model's total correctness over all classes. It is especially crucial in situations where the video recognition system's overall efficacy is at stake, like in surveillance or medical monitoring. Both precision and recall are important metrics to consider, particularly when dealing with imbalanced datasets that have underrepresented classes. Precision gauges how well the model can reduce false positives by calculating the accuracy of the predictions made in the positive class as shown in figure 2. The accuracy, precision and recall has reported more than 99.6%.

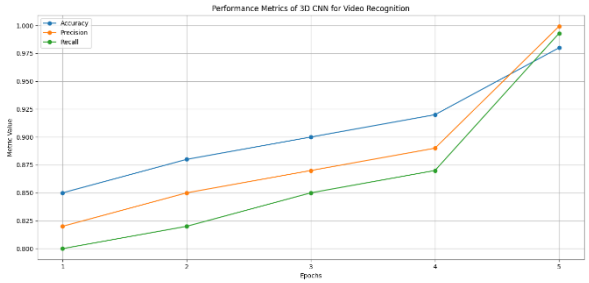

Figure 2: Accuracy, Precision, Recall for 3D- Video Recognition

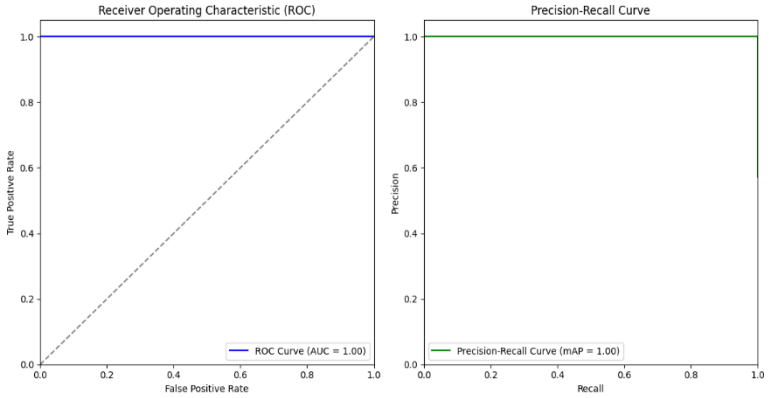

Figure 3: ROC and PR Curve for 3D- Video Recognition

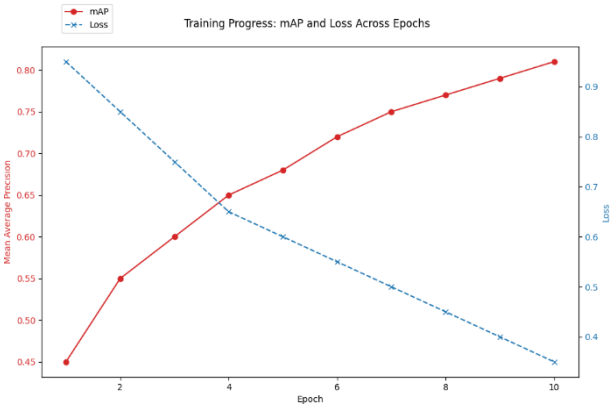

Figure 4: mAP and Loss Across Epoch for 3D- Video Recognition

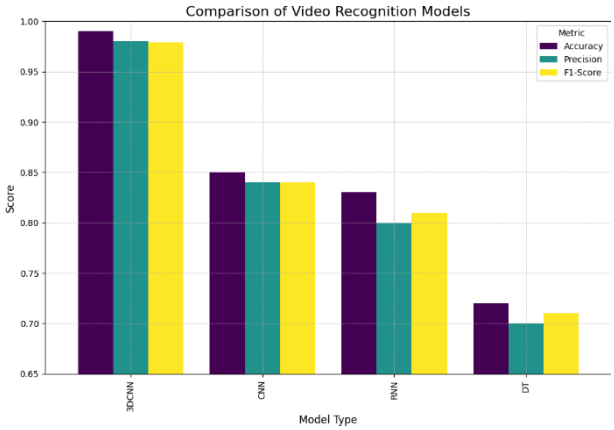

Figure 5: Video Recognition performance comparisons with various models

To evaluate the effectiveness of 3D Convolutional Neural Networks (3D-CNNs) in video recognition tasks, two essential assessment methods are the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves. The True Positive Rate (TPR) against False Positive Rate (FPR) at various threshold settings is plotted on the ROC curve, which offers information on how well the model discriminates across classes at various thresholds. An AUC that is near to one signifies exceptional model performance.Plotting Precision against Recall for various threshold values, the PR curve is especially helpful in situations where there is a large class imbalance. An increased area under the PR curve (AP) indicates that a greater percentage of positive samples are successfully retrieved by the model while retaining a high level of precision as in figure 3.

It is crucial to track mean Average Precision (mAP) and loss metrics over epochs when training 3D Convolutional Neural Networks (3D-CNNs) for video recognition in order to assess model performance and convergence. A complete measure of precision across different recall levels, the mAP metric is computed for each epoch and is especially significant when processing numerous classes in video data. Greater overall model accuracy in differentiating between various actions or events within the films is indicated by a higher mAP value.On the other hand, the loss metric measures how effectively the model predicts the labels; lower values correspond to greater performance. Generally, the loss should go down as the epochs go on, indicating that the model is learning efficiently as shown in figure 4.

Particularly designed for handling spatiotemporal data are 3D-CNNs. 3D-CNNs are superior at collecting motion information in videos because they expand conventional CNNs to incorporate the temporal dimension directly in the convolutional layers. Their capacity to comprehend temporal

dynamics makes them ideal for activities where understanding is critical, such as action identification or event detection. CNNs are good at recognizing spatial features, but they need other processes to deal with the time part of video input, such as optical flow or temporal modelling. Without integration with CNNs, RNNs especially those with LSTM or GRU configurations may have difficulty interpreting spatial features, but they are excellent at capturing temporal connections.

On the other hand, Decision Trees and their ensembles lack the capacity to process raw video data efficiently without extensive feature engineering. They typically underperform neural network approaches due to their inability to capture complex patterns in high-dimensional data. In performance metrics like accuracy, precision, recall, and F1-score, 3D-CNNs often outshine the other models in comprehensive video analysis tasks due to their integrated approach to learning both spatial and temporal features. However, the selection of the most appropriate model depends on factors such as computational resources, latency requirements, and the specific characteristics of the video data and task at hand as shown in figure 5.

5. Conclusion

In summary, 3D Convolutional Neural Networks (3D-CNNs) have become highly effective tools for video recognition applications due to their special capacity to extract both temporal and spatial characteristics directly from video input. 3D-CNNs are superior to regular CNNs at comprehending motion and dynamic patterns in videos because they extend CNNs to include the temporal dimension. They are especially well-suited for jobs like action recognition, gesture recognition, and activity detection in videos because of these capabilities. 3D-CNNs have proven to perform very well in a variety of fields, including as autonomous vehicles, healthcare, entertainment, and surveillance, by utilizing deep learning techniques. There are still issues, though, like dataset size, domain-specific variances, and computational complexity. To further develop the capabilities of 3D-CNNs for video, future research topics might concentrate on optimizing architectures, increasing efficiency, and improving interpretability.

References

[1] Kopuklu, Okan, Neslihan Kose, Ahmet Gunduz, and Gerhard Rigoll. "Resource efficient 3d convolutional neural networks." In Proceedings of the IEEE/CVF international conference on computer vision workshops, pp. 0-0. 2019.
[2] Pokkuluri, Kiran Sree, and SSSN Usha Devi Nedunuri. "A novel cellular automata classifier for covid-19 prediction." Journal of Health Sciences 10, no. 1 (2020): 34-38.
[3] Yang, Hao, Chunfeng Yuan, Bing Li, Yang Du, Junliang Xing, Weiming Hu, and Stephen J. Maybank. "Asymmetric 3d convolutional neural networks for action recognition." Pattern recognition 85 (2019): 1-12.
[4] Pokkuluri, Kiran Sree, SSSN Usha Devi Nedunuri, and Usha Devi. "Crop Disease Prediction with Convolution Neural Network (CNN) Augmented With Cellular Automata." Int. Arab J. Inf. Technol. 19, no. 5 (2022): 765-773.
[5] Maqsood, Ramna, Usama Ijaz Bajwa, Gulshan Saleem, Rana Hammad Raza, and Muhammad Waqas Anwar. "Anomaly recognition from surveillance videos using 3D convolution neural network." Multimedia Tools and Applications 80, no. 12 (2021): 18693-18716.
[6] Pokkuluri, Kiran Sree, and Devi Nedunuri Usha. "A secure cellular automata integrated deep learning mechanism for health informatics." Int. Arab J. Inf. Technol. 18, no. 6 (2021): 782-788.
[7] Sree, Pokkuluri Kiran, Phaneendra Varma Chintalapati, M. Prasad, Gurujukota Ramesh Babu, and PBV Raja Rao. "Waste Management Detection Using Deep Learning." In 2023 3rd International Conference on Computing and Information Technology (ICCIT), pp. 50-54. IEEE, 2023.
[8] Liu, Jiaqi, Zhenghao Li, Yongliang Tang, Wei Hu, and Jun Wu. "3D Convolutional Neural Network based on memristor for video recognition." Pattern Recognition Letters 130 (2020): 116-124.
[9] Rao, N. Raghava, Sree Pokkuluri Kiran, Tamboli Amena, A. Senthilkumar, R. Sivakumar, M. Ashok Kumar, and Sampathkumar Velusamy. "Enhancing rainwater harvesting and groundwater recharge efficiency with multi-dimensional LSTM and clonal selection algorithm." Groundwater for Sustainable Development (2024): 101167.
[10] Funke, Isabel, Sebastian Bodenstedt, Florian Oehme, Felix von Bechtolsheim, Jürgen Weitz, and Stefanie Speidel. "Using 3D convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition

in video." In International conference on medical image computing and computer-assisted intervention, pp. 467-475. Cham: Springer International Publishing, 2019.
[11] Lin, Beibei, Shunli Zhang, and Feng Bao. "Gait recognition with multiple-temporal-scale 3d convolutional neural network." In Proceedings of the 28th ACM international conference on multimedia, pp. 3054-3062. 2020.
[12] Liu, Jiawei, Zheng-Jun Zha, Xuejin Chen, Zilei Wang, and Yongdong Zhang. "Dense 3D-convolutional neural network for person re-identification in videos." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 15, no. 1s (2019): 1-19.
[13] Chintalapati, Phaneendra Varma, Gurujukota Ramesh Babu, Pokkuluri Kiran Sree, Satish Kumar Kode, and Gottala Surendra Kumar. "Usage of AI Techniques for Cyberthreat Security System in Android Mobile Devices." In International Conference On Innovative Computing And Communication, pp. 443-454. Singapore: Springer Nature Singapore, 2023.
[14] Accattoli, Simone, Paolo Sernani, Nicola Falcionelli, Dagmawi Neway Mekuria, and Aldo Franco Dragoni. "Violence detection in videos by combining 3D convolutional neural networks and support vector machines." Applied Artificial Intelligence 34, no. 4 (2020): 329-344.
[15] Pokkuluri, Kiran Sree, and Alex Khang. "Integration of Machine Learning Augmented With Biosensors for Enhanced Water Quality Monitoring." In Agriculture and Aquaculture Applications of Biosensors and Bioelectronics, pp. 181-192. IGI Global, 2024.
[16] Funke, Isabel, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel. "Video-based surgical skill assessment using 3D convolutional neural networks." International journal of computer assisted radiology and surgery 14 (2019): 1217-1225.
[17] Li, Jianing, Shiliang Zhang, and Tiejun Huang. "Multi-scale 3d convolution network for video based person re-identification." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 8618-8625. 2019.
[18] Li, Jing, Yandan Wang, John See, and Wenbin Liu. "Micro-expression recognition based on 3D flow convolutional neural network." Pattern Analysis and Applications 22 (2019): 1331-1339.
[19] Sree, Pokkuluri Kiran, Gurujukota Ramesh Babu, PBV Raja Rao, Phaneendra Varma Chintalapati, and M. Prasad. "Fake News Detection using Cellular Automata Based Deep Learning." In 2023 3rd International Conference on Computing and Information Technology (ICCIT), pp. 167-171. IEEE, 2023.
[20] Haddad, Jad, Olivier Lézoray, and Philippe Hamel. "3d-cnn for facial emotion recognition in videos." In Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5–7, 2020, Proceedings, Part II 15, pp. 298-309. Springer International Publishing, 2020.
[21] Reddy, Sai Prasanna Teja, Surya Teja Karri, Shiv Ram Dubey, and Snehasis Mukherjee. "Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks." In 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1-8. IEEE, 2019.
[22] Salama, Elham S., Reda A. El-Khoribi, Mahmoud E. Shoman, and Mohamed A. Wahby Shalaby. "A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition." Egyptian Informatics Journal 22, no. 2 (2021): 167-176.
[23] Pokkuluri, Kiran Sree, Usha Devi NSSSN, Martin Margala, and Prasun Chakrabarti. "Enhancing Image Segmentation Accuracy using Deep Learning Techniques." Journal of Advanced Research in Applied Sciences and Engineering Technology 49, no. 1 (2025): 139-148.
[24] Sharma, Shikhar, and Krishan Kumar. "ASL-3DCNN: American sign language recognition technique using 3-D convolutional neural networks." Multimedia Tools and Applications 80, no. 17 (2021): 26319-26331.
[25] Solovyev, Roman, Alexandr A. Kalinin, and Tatiana Gabruseva. "3D convolutional neural networks for stalled brain capillary detection." Computers in biology and medicine 141 (2022): 105089.
[26] Sree, Pokkuluri Kiran, Prasun Chakrabarti, Martin Margala, Phaneendra Varma Chintalapati, Gurujukota Ramesh Babu, and S. S. S. N. Usha Devi N. "Hybrid Cellular Automata with CNN for the Prediction of Secondary Structure of Protein." In *International Conference on Innovations in Data Analytics*, pp. 303-311. Singapore: Springer Nature Singapore, 2023.