

ORIGINAL

Investigating the Influence of Convolutional Operations on LSTM Networks in Video Classification

Investigación de la influencia de las operaciones convolucionales en las redes LSTM para la clasificación de vídeo

Manal Benzyane¹, Mourade Azroul¹, Imad Zeroual¹, Said Agoujl²

¹STI, IDMS, FST Errachidia, Moulay Ismail University of Meknes, Morocco.

²MMIS, MAIS, FST Errachidia, Moulay Ismail University, Meknes, Morocco.

Cite as: Benzyane M, Azroul M, Zeroual I, Agoujl S. Investigating the Influence of Convolutional Operations on LSTM Networks in Video Classification. Data and Metadata. 2023;2:152. <https://doi.org/10.56294/dm2023152>


Submitted: 10-08-2023

Revised: 21-09-2023

Accepted: 29-12-2023

Published: 30-12-2023

Editor: Prof. Dr. Javier González Argote 

Guest Editor: Yousef Farhaoui 

Note: Paper presented at the International Conference on Artificial Intelligence and Smart Environments (ICAISE'2023).

ABSTRACT

Video classification holds a foundational position in the realm of computer vision, involving the categorization and labeling of videos based on their content. Its significance resonates across various applications, including video surveillance, content recommendation, action recognition, video indexing, and more. The primary objective of video classification is to automatically analyze and comprehend the visual information embedded in videos, facilitating the efficient organization, retrieval, and interpretation of extensive video collections. The integration of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks has brought about a revolution in video classification. This fusion effectively captures both spatial and temporal dependencies within video sequences, leveraging the strengths of CNNs in extracting spatial features and LSTMs in modeling sequential and temporal information. ConvLSTM and LRCN (Long-term Recurrent Convolutional Networks) are two widely embraced architectures that embody this fusion. This paper seeks to investigate the impact of convolutions on LSTM networks in the context of video classification, aiming to compare the performance of ConvLSTM and LRCN.

Keywords: Video Classification; Convolution; LSTM; ConvLSTM; LRCN.

RESUMEN

La clasificación de vídeos ocupa un lugar fundamental en el ámbito de la visión informática, ya que consiste en categorizar y etiquetar vídeos en función de su contenido. Su importancia se extiende a diversas aplicaciones, como la videovigilancia, la recomendación de contenidos, el reconocimiento de acciones y la indexación de vídeos, entre otras. El objetivo principal de la clasificación de vídeos es analizar y comprender automáticamente la información visual contenida en los vídeos, facilitando la organización, recuperación e interpretación eficientes de extensas colecciones de vídeos. La integración de redes neuronales convolucionales (CNN) y redes de memoria a corto plazo (LSTM) ha supuesto una revolución en la clasificación de vídeos. Esta fusión captura eficazmente las dependencias espaciales y temporales dentro de las secuencias de vídeo, aprovechando los puntos fuertes de las CNN en la extracción de características espaciales y de las LSTM en el modelado de la información secuencial y temporal. ConvLSTM y LRCN (Long-term Recurrent Convolutional Networks) son dos arquitecturas ampliamente aceptadas que incorporan esta fusión. Este artículo pretende investigar el impacto de las convoluciones en las redes LSTM en el contexto de la clasificación de vídeo, con el objetivo de comparar el rendimiento de ConvLSTM y LRCN.

Palabras clave: Clasificación de vídeo; Convolución; LSTM; ConvLSTM; LRCN.

INTRODUCTION

Video classification is a crucial undertaking in computer vision, entailing video categorization based on content. Deep learning has marked significant progress in this domain, with ConvLSTM and LRCN emerging as prominent architectural advancements. The integration of Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs), observed in ConvLSTM and LRCN, has exerted a profound influence on video classification. ConvLSTM extends LSTM functionality by incorporating convolutional operations within its cells, enabling direct learning of spatial representations from video sequences and modeling of long-term temporal dependencies. ConvLSTM's convolutional operations adeptly extract spatial features, capturing frame-level patterns and object information, and enabling recognition of complex actions.⁽¹⁾

In contrast, LRCN adopts a sequential approach, utilizing CNNs initially to capture frame-level characteristics encoding spatial information. Subsequently, these features undergo processing in an LSTM network to discern temporal dependencies between frames. LRCN strategically leverages CNNs for spatial details and LSTMs for temporal context, enabling a comprehensive understanding of movement patterns and temporal nuances.⁽²⁾

Following this introduction, the subsequent sections of this paper are organized as follows. Section 2 provides an overview of the related works, and in section 3 we provide an overview of the methodology employed, including a description of the datasets used. Additionally, it presents the theoretical background of key architectures such as CNN, LSTM, ConvLSTM, and LRCN. In Section 4, we present the main findings obtained from the classifiers' performance evaluation using different datasets. Finally, Section 5 wraps up the paper by providing a summary of the key findings and offering perspectives for future research endeavors.

Related Works

In the realm of video classification, several noteworthy works have contributed to advancing the field. Researchers have explored diverse methodologies, techniques, and architectures to enhance the accuracy and efficiency of video classification systems. One influential approach involves leveraging deep learning, particularly Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs).

Studies such as various studies^(3,4,5) delved into the fusion of CNNs and LSTMs, exemplified by architectures like ConvLSTM. These architectures have demonstrated notable success in capturing both spatial and temporal dependencies within video sequences, thereby improving the overall performance of video classification tasks.

Continuing the exploration of innovative architectures, the Long-term Recurrent Convolutional Network (LRCN) has emerged as a prominent example in the fusion of CNNs and LSTMs, as demonstrated in the work by several studies.^(4,5,6,7) LRCN follows a sequential approach, initially utilizing CNNs to capture frame-level characteristics encoding spatial information. These features are then fed into an LSTM network to capture temporal dependencies between frames. LRCN's strategic combination of CNNs for spatial details and LSTMs for temporal context enables it to understand movement patterns and temporal nuances. The success of LRCN further reinforces the effectiveness of combining spatial and temporal features for robust video classification.

These related works collectively underscore the dynamic and evolving landscape of video classification research, with a focus on integrating deep learning, exploring novel features, and leveraging transfer learning for improved performance and versatility.

METHODS

Dataset

In this study, we employed three datasets to conduct our experiments: UCF11, UCF50, and DynTex.

UCF11: The UCF11 dataset comprises a total of 1600 video clips, meticulously categorized into 11 diverse action types.⁽⁸⁾ Recognized as a benchmark dataset within the field of video classification, UCF11 is widely utilized to evaluate algorithms designed for the recognition and categorization of dynamic activities. Its extensive content encapsulates a broad array of human endeavors, ranging from sports activities such as basketball, biking, diving, golf swing, horse riding, and swinging, to everyday actions like tennis swing, forming a rich and varied collection.⁽⁹⁾

Renowned for its role as a standard reference, UCF11 serves as a crucial resource for assessing the efficacy of algorithms in handling a wide spectrum of dynamic activities.

UCF50: The UCF50 dataset stands as a prominent benchmark in the realm of video classification, offering a comprehensive collection that spans 50 diverse action categories. Comprising a total of 6618 realistic videos sourced from YouTube,⁽⁸⁾ this dataset presents a rich and varied set of activities for analysis. The action categories within UCF50 encapsulate a broad spectrum, featuring activities such as applying eye makeup, applying lipstick, band marching, archery, baby crawling, balance beam, basketball dunk, baseball pitch,

biking, bench press, billiards, bowling, and more. Renowned for its expansive and realistic content, UCF50 serves as a valuable resource for assessing and advancing video classification algorithms across a diverse array of real-world scenarios.

DynTex: The DynTex dataset stands out as a meticulously curated collection of video sequences centered around dynamic tissues, comprising a total of 522 videos showcasing dynamic textures.⁽¹⁰⁾ These videos are thoughtfully organized across five distinct dynamic textures: Clouds-Steam, Flags, Fire, Water, and Trees, presenting a diverse array of visual patterns and movements. This dataset plays a crucial role in the field of computer vision, serving as a cornerstone for dynamic texture-related tasks such as analysis, classification, and recognition. Widely employed by researchers and developers, the DynTex dataset proves to be a valuable resource for the in-depth study and advancement of algorithms aimed at proficiently capturing and analyzing the temporal intricacies inherent in dynamic textures within videos.

Convolutional neural networks

Deep learning models, specifically convolutional neural networks (CNNs), represent a ubiquitous and highly impactful approach in the realm of computer vision tasks. Renowned for their prowess in extracting meaningful information from visual data, CNN architectures excel in various image-related applications, including classification, detection, and recognition. The intrinsic design of CNNs encompasses convolutional layers, where filters are applied for intricate processing of input images, pooling layers that effectively decrease spatial dimensions, and fully connected layers tailored for tasks such as classification or regression.^(11,12,13,14,15,16,17,18,19,20,21) This versatile architecture has become instrumental in addressing complex visual challenges and remains at the forefront of advancements in computer vision re-search.

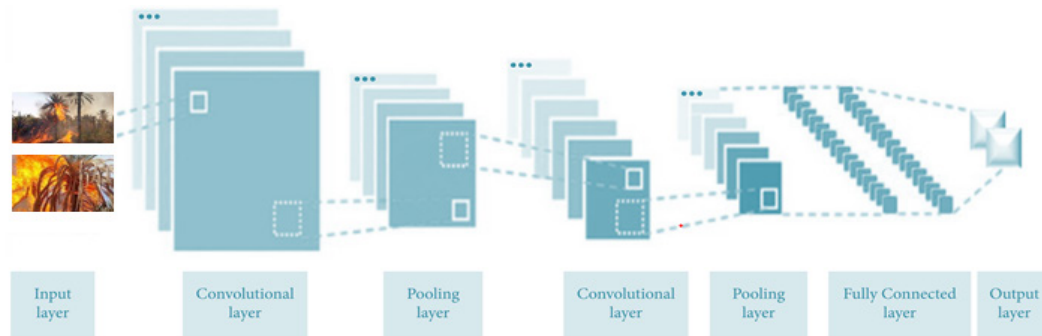


Figure 1. Architectures of convolutional neural networks

In the dynamic field of video classification, Convolutional Neural Networks (CNNs) showcase their effectiveness by proficiently capturing both spatial and temporal dimensions of visual data. Operating on a per-frame basis within a video, CNNs intricately extract frame-level features, enabling the discernment of complex visual patterns unique to each frame.

This spatial understanding forms the foundation for robust video analysis. To complement this spatial comprehension with temporal context, recurrent neural networks (RNNs) and Long Short-Term Memory networks (LSTMs) are frequently integrated. These networks specialize in processing sequential frames, facilitating the capture of temporal dependencies inherent in video sequences. RNNs and LSTMs contribute a crucial temporal perspective, allowing the model to grasp the evolving dynamics and relationships between consecutive frames over time.

This fusion of CNNs with specialized temporal processing networks exemplifies a comprehensive approach to video classification, ensuring that the model not only discerns intricate spatial details but also comprehends the nuanced temporal evolution within video content. It stands as a testament to the sophisticated strategies employed to enhance the discernment capabilities of video classification systems in the ever-evolving landscape of computer vision.

Long Short-Term Memory

Long Short-Term Memory (LSTM) represents a specialized class of recurrent neural networks (RNNs), meticulously crafted to overcome the vanishing gradient problem that often impedes traditional RNNs. Unlike its predecessors, LSTM incorporates a memory cell and gating mechanisms, allowing it to expertly capture extended, long-term dependencies within sequential data. This architectural refinement proves particularly advantageous in tasks demanding a nuanced understanding of temporal relationships, such as speech recognition and video analysis, where the ability to discern intricate patterns over time is paramount.⁽¹⁰⁾

The distinguishing feature of LSTM lies in its intricate chained structure, vividly depicted in figure 2. This

structure facilitates a sequential flow of information within the network, empowering LSTM to systematically capture and retain relevant con-textual information over time. At each step along the chain, the model engages in the retention and updating of pertinent information, ensuring a continuous and adaptive understanding of the evolving context within the sequential data.

This nuanced architecture has propelled LSTM to widespread adoption across diverse domains. Its proficiency in handling sequential data, coupled with its capacity to grasp long-term dependencies, positions LSTM as a versatile and powerful choice. The model's effectiveness in retaining context over extended sequences has made it instrumental in enhancing the capabilities of systems engaged in tasks requiring intricate temporal analysis, solidifying its role as a cornerstone in advanced data sequence processing.

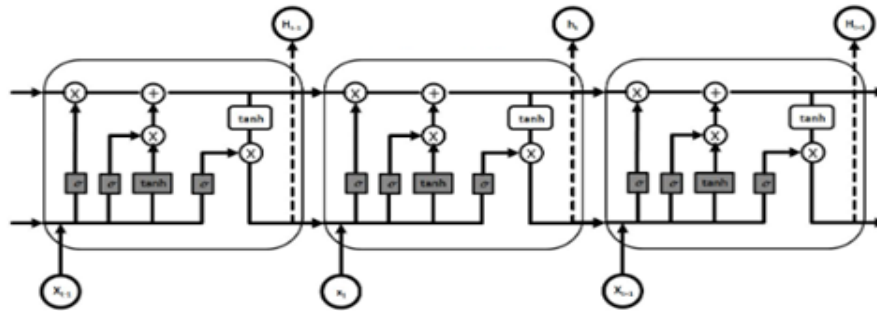


Figure 2. Structure of an LSTM network⁽¹⁴⁾

Convolutional Long Short-Term Memory (ConvLSTM)

Convolutional Long Short-Term Memory (ConvLSTM) represents a sophisticated evolution from the conventional Long Short-Term Memory (LSTM) architecture, introducing a transformative modification to bolster its capabilities in handling sequential data with spatial dependencies. The fundamental shift lies in the adaptation of the LSTM module, where fully-connected gates are replaced with convolutional gates.

The essence of ConvLSTM's innovation lies in its distinctive approach to spatial information integration and the capturing of spatial dependencies within sequential data, especially pertinent in domains like video analysis and spatiotemporal sequences. This is achieved through the strategic use of convolution operations at each gate, a departure from the matrix multiplication employed by traditional LSTMs.

This architectural modification enables ConvLSTM to capitalize on the strengths of convolutional neural networks (CNNs), leveraging local receptive fields and shared weights. By doing so, ConvLSTM gains an enhanced ability to learn intricate spatial representations and effectively capture long-term dependencies present in the data.⁽¹⁵⁾ This nuanced approach positions ConvLSTM as a robust solution in scenarios where understanding spatial context is crucial.

The utilization of convolutional gates in ConvLSTM represents a paradigm shift, providing a unique advantage in tasks demanding a sophisticated interplay between temporal and spatial dimensions within sequential data. This refined architecture elevates ConvLSTM as a formidable tool, showcasing its utility in cutting-edge applications that require an in-depth comprehension of spatial relationships in evolving data sequences.

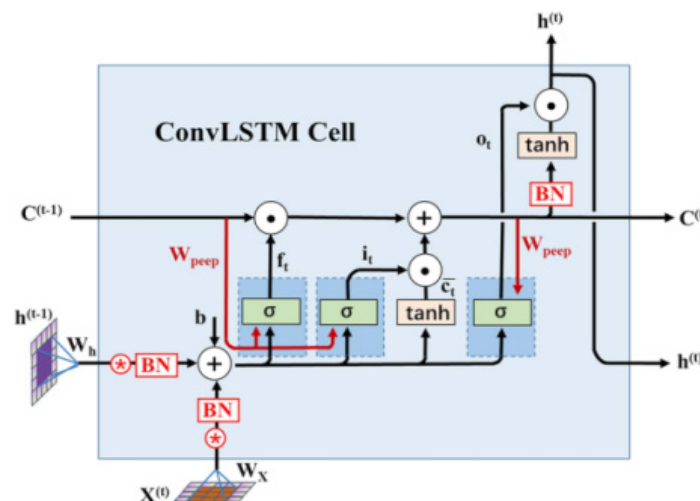


Figure 3. ConvLSTM cell architecture⁽¹⁶⁾

Long-term Recurrent Convolutional Networks (LRCN)

Long-term Recurrent Convolutional Networks (LRCN) adopt a meticulous sequential strategy for video classification, demonstrating a comprehensive and nuanced approach to decoding dynamic visual content. The process is initiated by harnessing the power of Convolutional Neural Networks (CNNs) to extract features at the frame level from each video frame. The meticulously extracted features undergo a sequential analysis within a Long Short-Term Memory (LSTM) network.

This strategic integration of LSTM serves as a pivotal step, enabling LRCN to model and comprehend intricate temporal dependencies existing between successive frames in a video sequence. The sequential scrutiny of frame-level features unfolds a dynamic narrative, allowing LRCN to effectively capture evolving dynamics and con-textual nuances present in video sequences.

LRCN's sequential methodology extends beyond recognizing individual frame patterns; it involves discerning nuanced relationships and narrative evolution be-tween frames over time. This detailed understanding empowers LRCN to effectively capture dynamic and contextual information within video sequences, significantly augmenting its proficiency in classifying and comprehending video content.⁽¹⁷⁾ This sophisticated methodology positions LRCN as a potent tool for tasks demanding a holistic grasp of both spatial and temporal dimensions within dynamic visual data.

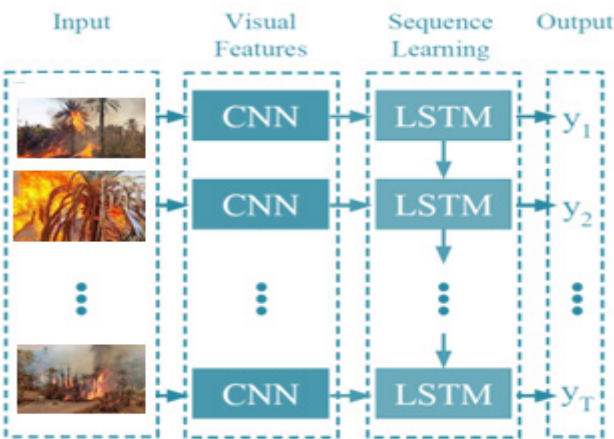


Figure 4. Architectures of LRCN

RESULTS

Our research involved a meticulously structured series of experiments employing ConvLSTM and LRCN models, with a deliberate focus on understanding their performance dynamics. We initiated our analysis by closely examining the UCF11 dataset and systematically expanded our investigations to encompass the broader contexts presented by the UCF50 and DynTex datasets. Through an in-depth comparative analysis of ConvLSTM and LRCN models across these diverse datasets, our study sought to unravel nuanced insights into the intricate interplay of convolutional operations on Long Short-Term Memory (LSTM) networks within the domain of video classification.

The core of our investigation delved into the simultaneous integration of Convolutional Neural Networks (CNNs) and LSTM networks. Our scrutiny extended to scenarios where CNNs and LSTMs collaboratively contribute to the video classification process, enabling us to discern the combined effects of these architectural components. Furthermore, our study rigorously examined the implications of adopting CNNs as the initial step to extract frame-level features from individual video frames, followed by the application of LSTMs.

This multifaceted experimental design was crafted to systematically dissect and comprehend the intricate dynamics surrounding convolutional influences on LSTM networks. By adopting this comprehensive approach, we aimed to provide a detailed understanding of the multifaceted impact of convolutional operations on the com-plex landscape of video classification tasks. The table below illustrates the obtained results.

Table 1. Comparison of ConvLSTM and LRCN on data			
Method	DynTex	UCF11	UCF50
ConvLSTM	0,56	0,62	0,79
LRCN	0,71	0,77	0,93

For the DynTex dataset, the ConvLSTM model exhibited an accuracy of 0,56, whereas the LRCN model demonstrated superior performance with a higher accuracy of 0,71. Transitioning to the UCF11 dataset, the

ConvLSTM model achieved an accuracy of 0,62, while the LRCN model excelled further with a higher accuracy of 0,77. In the UCF50 dataset, the ConvLSTM model attained an accuracy of 0,79, yet the LRCN model outshone with an even higher accuracy of 0,93. These findings underscore a consistent trend wherein the LRCN model consistently outperformed the ConvLSTM model across all three datasets, emphasizing its superior efficacy in video classification tasks.

CONCLUSION

In conclusion, video classification stands as a crucial domain within computer vision and multimedia content interpretation, presenting both challenges and avenues for the development of robust classification models. Our investigation delved into the impact of convolutions on LSTM networks for video classification, revealing that the sequential utilization of CNNs for initial feature extraction, followed by LSTMs, yields superior results compared to the simultaneous use of CNNs and LSTMs. This sequential approach demonstrated enhanced performance and accuracy.

While our pilot study provides valuable insights, it also highlights opportunities for refinement in video classification methodologies. Future research endeavors could focus on exploring diverse CNN and LSTM architectures, optimizing their parameters to further enhance the efficacy of the sequential approach. By contributing to ongoing advancements in this field, such efforts hold the potential to elevate the precision and capabilities of video classification models.

REFERENCES

1. Z. Sun et M. Zhao, « Short-Term Wind Power Forecasting Based on VMD Decomposition, ConvLSTM Networks and Error Analysis », IEEE Access, vol. 8, p. 134422-134434, 2020, doi: 10.1109/ACCESS.2020.3011060.
2. M. S. Uzzaman, C. Debnath, D. M. A. Uddin, M. M. Islam, M. A. Talukder, et S. Parvez, « LRCN Based Human Activity Recognition from Video Data ». Rochester, NY, 25 août 2022. doi: 10.2139/ssrn.4173741.
3. M. Alharbi, S. K. Rajagopal, S. Rajendran, et M. Alshahrani, « Plant Disease Classification Based on ConvLSTM U-Net with Fully Connected Convolutional Layers », TS, vol. 40, no 1, p. 157-166, févr. 2023, doi: 10.18280/ts.400114.
4. U. Singh et N. Singhal, « Exploiting Video Classification Using Deep Learning Models for Human Activity Recognition », in Computer Vision and Robotics, P. K. Shukla, K. P. Singh, A. K. Tripathi, et A. Engelbrecht, Éd., in Algorithms for Intelligent Systems. Singapore: Springer Nature, 2023, p. 169-179. doi: 10.1007/978-981-19-7892-0_14.
5. W.-Y. Wang, H.-C. Li, Y.-J. Deng, L.-Y. Shao, X.-Q. Lu, et Q. Du, « Generative Adversarial Capsule Network With ConvLSTM for Hyperspectral Image Classification », IEEE Geoscience and Remote Sensing Letters, vol. 18, no 3, p. 523-527, mars 2021, doi: 10.1109/LGRS.2020.2976482.
6. Y. Tang, J. Huang, et S. Gao, « Research on Fault Classification Model of TE Chemical Process Based on LRCN », in 2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI), juill. 2021, p. 118-122. doi: 10.1109/DTPI52967.2021.9540171.
7. S. Gogineni, G. Suryanarayana, et K. L. S. Soujanya, « Pruning Long-term Recurrent Convolutional Networks for Video Classification and captioning », in 2020 International Conference on Smart Electronics and Communication (ICOSEC), sept. 2020, p. 215-221. doi: 10.1109/ICOSEC49089.2020.9215414.
8. S. Zebhi, S. M. T. AlModarresi, et V. Abootalebi, « Action Recognition in Videos Using Global Descriptors and Pre-trained Deep Learning Architecture », in 2020 28th Iranian Conference on Electrical Engineering (ICEE), Tabriz, Iran: IEEE, août 2020, p. 1-4. doi: 10.1109/ICEE50131.2020.9261038.
9. Y. Cheng, Y. Yang, H.-B. Chen, N. Wong, et H. Yu, « S3-Net: A Fast and Lightweight Video Scene Understanding Network by Single-shot Segmentation », in 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Wai-koloa, HI, USA: IEEE, janv. 2021, p. 3328-3336. doi: 10.1109/WACV48630.2021.00337.
10. M. Benzyane, I. Zeroual, M. Azrou, et S. Agoujil, « Convolutional Long Short-Term Memory Network Model for Dynamic Texture Classification: A Case Study », in International Conference on Advanced Intelligent Systems for Sustainable Development, J. Kacprzyk, M. Ezziyyani, et V. E. Balas, Éd., in Lecture Notes in Networks and Systems. Cham: Springer Nature Switzerland, 2023, p. 383-395. doi: 10.1007/978-3-031-26384-2_33.

11. Y. LeCun, Y. Bengio, et G. Hinton, « Deep learning », Nature, vol. 521, no 7553, Art. no 7553, mai 2015, doi: 10.1038/nature14539.
12. T. J. Brinker et al., « Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review », Journal of Medical Internet Research, vol. 20, no 10, p. e11936, oct. 2018, doi: 10.2196/11936.
13. Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, et X. Xue, « Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification », in Pro-ceedings of the 23rd ACM international conference on Multimedia, Brisbane Aus-tralia: ACM, oct. 2015, p. 461-470. doi: 10.1145/2733373.2806222.
14. K. Luan et T. Matsumaru, « Dynamic Hand Gesture Recognition for Robot Arm Teaching based on Improved LRCN Model », in 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO), déc. 2019, p. 1269-1274. doi: 10.1109/ROBIO49542.2019.8961787.
15. W. Ye, J. Cheng, F. Yang, et Y. Xu, « Two-Stream Convolutional Network for Improving Activity Recognition Using Convolutional Long Short-Term Memory Networks », IEEE Access, vol. 7, p. 67772-67780, 2019, doi: 10.1109/ACCESS.2019.2918808.
16. H. Sun, Y. Yang, Y. Chen, X. Liu, et J. Wang, « Tourism demand forecasting of multi-attractions with spatiotemporal grid: a convolutional block attention module model », Information Technology & Tourism, p. 1-29, avr. 2023, doi: 10.1007/s40558-023-00247-y.
17. J. Choi, J. S. Lee, M. Ryu, G. Hwang, G. Hwang, et S. J. Lee, « Attention-LRCN: Long-term Recurrent Convolutional Network for Stress Detection from Photoplethysmography », in 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA), juin 2022, p. 1-6. doi: 10.1109/MeMeA54994.2022.9856417.
18. Romero-Carazas R. Prompt lawyer: a challenge in the face of the integration of artificial intelligence and law. Gamification and Augmented Reality 2023;1:7-7. <https://doi.org/10.56294/gr20237>.
19. Gonzalez-Argote D, Gonzalez-Argote J, Machuca-Contreras F. Blockchain in the health sector: a systematic literature review of success cases. Gamification and Augmented Reality 2023;1:6-6. <https://doi.org/10.56294/gr20236>.
20. Tarik, A., and all. "Recommender System for Orientation Student" Lecture Notes in Networks and Systems, 2020, 81, pp. 367-370. https://doi.org/10.1007/978-3-030-23672-4_27
21. Gonzalez-Argote J. Analyzing the Trends and Impact of Health Policy Research: A Bibliometric Study. Health Leadership and Quality of Life 2023;2:28-28. <https://doi.org/10.56294/hl202328>.
22. Gonzalez-Argote J. A Bibliometric Analysis of the Studies in Modeling and Simulation: Insights from Scopus. Gamification and Augmented Reality 2023;1:5-5. <https://doi.org/10.56294/gr20235>.
23. Sossi Alaoui, S., and all. "A comparative study of the four well-known classi-fication algorithms in data mining", Lecture Notes in Networks and Systems, 2018, 25, pp. 362-373. https://doi.org/10.1007/978-3-319-69137-4_32

FINANCING

No financing.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Manal Benzyane, Mourade Azrou, Imad Zeroual, and Said Agoujl.

Research: Manal Benzyane, Mourade Azrou, Imad Zeroual, and Said Agoujl.

Drafting - original draft: Manal Benzyane, Mourade Azrou, Imad Zeroual, and Said Agoujl.

Writing - proofreading and editing: Manal Benzyane, Mourade Azrou, Imad Zeroual, and Said Agoujl.