Indonesian Journal of Electrical Engineering and Computer Science
Vol. 42, No. 1, April 2026, pp. 174~182
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v42.i1.pp174-182

Enhanced long-term recurrent convolutional network for video classification

Manal Benzyane[1], Mourade Azrour[1], Said Agoujil[2]
[1]IMIA, MSIA, Faculty of Sciences and Techniques, Moulay Ismail University of Meknes, Errachidia, Morocco
[2]IMIA, MSIA, ENCG, Moulay Ismail University of Meknes, Meknes, Morocco

Article Info

Article history:

Received Jul 7, 2025
Revised Feb 26, 2026
Accepted Mar 4, 2026

Keywords:

Convolutional neural network
DynTex
Long short-term memory
LRCN
UCF11
UCF50
Video classification

ABSTRACT

Video classification is essential in computer vision, enabling automated understanding of dynamic content in applications such as surveillance, autonomous systems, and content recommendation. Traditional long-term recurrent convolutional network (LRCN) models, however, often struggle to capture complex spatio-temporal patterns, limiting classification performance across diverse video datasets. To address this limitation, we propose an enhanced LRCN with architectural refinements, optimized filter sizes, and hyperparameter tuning, improving both temporal modeling and spatial feature extraction. Experimental results on three benchmark datasets DynTex, UCF11, and UCF50 demonstrate that the proposed model achieves accuracies of 0.90 on DynTex (+26.8% over standard LRCN), 0.92 on UCF11 (+19.5%), and 0.94 on UCF50 (+1.1%), consistently outperforming ConvLSTM, LRCN, and other state-of-the-art approaches. These findings indicate that the enhanced LRCN effectively captures spatial and temporal dynamics in video sequences, setting a new benchmark for video classification. The study highlights the impact of architectural innovation and parameter optimization, providing a solid foundation for future research on scalable and efficient deep learning models for dynamic content analysis.

Corresponding Author:

Manal Benzyane
IMIA, MSIA, Faculty of Sciences and Techniques, Moulay Ismail University of Meknes
Errachidia, Morocco
Email: m.benzyane@edu.umi.ac.ma

1. INTRODUCTION

Video classification is a pivotal task in computer vision [1], enabling the automatic interpretation of dynamic visual content for a variety of applications, ranging from human activity recognition [2] to autonomous systems, medical diagnosis [3], and video surveillance [4]. As video data typically involves both spatial and temporal dimensions, the challenge of effectively processing and understanding these sequences has become increasingly important. Traditional image classification techniques, which focus solely on spatial features, fall short when it comes to interpreting video data, as they fail to capture the temporal dependencies between frames that are crucial for understanding dynamic events. The complexity of video classification lies in effectively capturing both spatial and temporal information from video sequences. Spatial features, such as objects, textures, and environments, need to be extracted from each individual frame, while temporal features, such as motion and interactions between objects, need to be modeled across frames. To address these challenges, various deep learning models have been developed, and among these, long-term recurrent convolutional networks (LRCNs) have emerged as a prominent architecture for sequential data analysis. LRCNs combine the power of convolutional neural networks (CNNs) for the extraction of spatial features, with long short-term memory (LSTM) networks for the modeling of temporal dependencies [5], thus

enabling the model to handle both dimensions simultaneously. While LRCNs have achieved notable success in video classification tasks, the standard architecture faces limitations when dealing with complex datasets or dynamic actions that require more refined feature extraction capabilities. These challenges arise due to the inherent variability in video content, including changes in scene composition, object motion, and environmental context [6]. Furthermore, the standard LRCN architecture may struggle to extract finer, high-level features when confronted with large-scale, diverse datasets. this study proposes an enhanced LRCN model designed to improve temporal feature modeling and refine spatial feature extraction. Our approach introduces optimized filter configurations and architectural refinements to enhance the model's capacity for capturing fine-grained video dynamics. These modifications contribute to a more accurate and robust classification of complex video sequences. The significance of this study lies in its potential to advance video classification by addressing the shortcomings of existing LRCN architectures. An efficient and precise classification model is crucial for real-world applications such as intelligent surveillance, autonomous navigation, and medical video analysis. Our enhanced LRCN model is extensively evaluated on three benchmark datasets DynTex, UCF11, and UCF50 covering a diverse range of actions, textures, and complexities. Experimental results demonstrate that the proposed model outperforms the standard LRCN architecture, showcasing improved accuracy and robustness across different datasets. By introducing this refined architecture, we aim to contribute to the ongoing advancements in video classification, providing a more effective and scalable solution for analyzing complex video data. Our work lays the foundation for future research in video understanding and classification, offering new directions for optimizing deep-learning models in dynamic and large-scale video environments.

The organization of this paper is as follows: section 2 delivers a thorough review of the existing literature to contextualize the study. In section 3 details the methodology, covering the datasets used, the theoretical basis of standard LRCN architectures, and a full description of the proposed model. Section 4 presents the key results of evaluating our proposed model compared to the LRCN model, which outperformed the ConvLSTM model discussed in our previous study [7], and other state-of-the-art methods using various datasets. Finally, section 5 concludes the paper with a summary of the main findings and insights, along with suggestions for future research directions.

2. RELATED WORKS

Video classification has witnessed rapid advancements, driven by the need to analyze diverse video content across various domains. A wide array of methods has been proposed, each tailored to address specific challenges related to spatiotemporal information extraction, multimodal data fusion, or classification tasks. Bi-directional long short-term memory networks (BiLSTM) have been widely adopted for their ability to model long-term dependencies in sequential data. For example, the BiLSTM-multimodal attention fusion temporal classification (BiLSTM-MAFTC) integrates BiLSTM with spatial and channel attention mechanisms to fuse features from multiple modalities, capturing complementary information such as movement trajectories and positional data [8]. Similarly, convolutional approaches, like deep convolutional neural networks (DCNNs), have been explored for their capability to extract discriminative features from video content. DCNNs have been applied alongside recurrent models like gated recurrent units (GRU) and recurrent neural networks (RNNs) for categorizing video data based on textual metadata such as titles and tags [9]. Transformer-based architectures have emerged as powerful tools for video classification. For example, the multi-task video transformer network (MTVTNet) leverages the swin transformer architecture to concurrently detect and classify multiple activities, making it highly effective for analyzing dynamic video content such as construction site operations. Similarly, attention mechanisms and graph-based learning have been integrated into frameworks like MALL-CNN [10], which not only models label co-occurrences but also aggregates frame-level features into meaningful video-level representations for multi-label classification tasks. Hybrid approaches combining CNNs and RNNs have also gained traction. For instance, in livestock behavior analysis, 3D convolutional neural networks (C3D) are used to extract spatial features, which are then processed by convolutional long short-term memory (ConvLSTM) networks to capture temporal dependencies. This combination enhances the accuracy of behavior classification tasks. In environmental monitoring, CNN-LSTM architectures have been applied to classify wave heights in ocean videos [11], utilizing monoscopic video inputs and sequential modeling. Optimization-based ensemble methods have been proposed to improve classification performance in challenging tasks like detecting video authenticity. Weighted and evolving ensembles combining 3D CNNs and CNN-RNNs have been enhanced through particle swarm optimization (PSO) [12], which optimizes network topologies and hyperparameters. These approaches effectively balance spatial-temporal feature extraction and classification complexity. In medical video analysis, ResNet-based architectures, such as ResNet-50 and ResNet-101, are used for detailed classification of medical imaging videos [13]. Techniques like data augmentation and contrast enhancement have been adopted to enhance the robustness of these models for specialized tasks, such as lesion

classification. For traffic state classification, the interactive multiple model (IMM) filter offers a unique approach by combining extended Kalman filters with a multi-class macroscopic model. This method avoids traditional training phases and provides accurate state estimations and classifications, such as distinguishing between free-flow and congested traffic states [14]. These varied methodologies illustrate the breadth of video classification research, highlighting innovations that leverage advancements in deep learning, optimization techniques, and domain-specific adaptations to tackle challenges across diverse application areas. Despite these advancements, challenges persist in achieving high accuracy across diverse datasets and efficiently handling complex video sequences. Recent studies have explored optimizing the LRCN architecture through modifications in filter size, depth, and recurrent layers, aiming to improve its effectiveness in video classification tasks. This work builds on these advancements by proposing an enhanced LRCN model, addressing the limitations of the standard architecture, and demonstrating its improved performance on three benchmark datasets: DynTex, UCF11, and UCF50.

3. METHOD

3.1. Dataset

To evaluate the effectiveness of our proposed model, we conducted experiments on three widely recognized and diverse datasets, each designed to challenge different aspects of video classification performance:

- DynTex: The dynamic texture dataset comprises videos that capture natural scenes with dynamic patterns [15], such as flowing water, waving foliage, and flickering flames [15]. These videos emphasize the temporal aspect of motion and require models to effectively capture fine-grained temporal dynamics. The dataset is ideal for evaluating the capability of video classification models to recognize subtle and continuous patterns across frames.
- UCF11: Also known as the YouTube action dataset, UCF11 contains 11 categories of human activities [16] such as biking, diving, and walking. The videos in this dataset feature diverse environments [17], camera angles, and action speeds, providing a balanced mix of spatial and temporal challenges. It serves as a benchmark for models to classify human actions with moderate complexity.
- UCF50: This is a larger and more complex dataset compared to UCF11, featuring 50 action categories [18] ranging from athletic activities like basketball and soccer to everyday actions such as brushing teeth and playing guitar. The dataset includes significant variations in lighting conditions, background clutter, camera motion, and action dynamics. These characteristics make UCF50 a rigorous benchmark for evaluating a model's ability to generalize across diverse and challenging scenarios.

Each dataset was selected to test the model's performance at varying levels of complexity, from recognizing subtle temporal patterns in dynamic textures to identifying intricate human activities under real-world conditions. By using these datasets, we ensure a comprehensive evaluation of the proposed model, highlighting its robustness and adaptability across different video classification tasks.

The datasets were preprocessed to standardize video resolution and frame rates, ensuring compatibility with the input requirements of our model. For each dataset, we split the data into training, validation, and testing sets to rigorously evaluate the model's learning ability, generalization performance, and resistance to overfitting. The diversity in these datasets allows for a detailed assessment of the model's strengths and limitations in video classification.

3.2. Long-term recurrent convolutional networks

The LRCN is a deep learning architecture that takes a sequential approach to video classification, effectively integrating spatial and temporal analysis to interpret dynamic video content. LRCN combines two powerful components: CNNs and LSTM networks. Initially, CNNs are employed to extract spatial features from individual video frames [19]. CNNs are a class of neural networks specifically designed to recognize visual patterns such as shapes, textures, edges, and objects by applying convolutional filters across the input image. These spatial features provide critical information about the content and structure of each frame, forming the foundation for subsequent temporal analysis. Unlike traditional methods that treat frames independently, the extracted spatial features are then passed to an LSTM network [20], LSTM is a type of RNN capable of modeling long-term temporal dependencies. It uses memory cells and gating mechanisms to selectively retain or forget information over time, making it particularly suitable for sequential data such as videos. This enables the LRCN to capture how visual information evolves across frames, including motion patterns [20], object interactions [21], and sequential events [22]. Figure 1 illustrates the LRCN architecture, showing the sequential flow from frame-wise CNN feature extraction to LSTM-based temporal modeling.

In recent advancements in video classification, architectures that combine convolutional layers with sequential processing, like the LRCN, have demonstrated significant promise. Following our previous study

[7], which showed that LRCN outperformed ConvLSTM in terms of both accuracy and robustness for video classification tasks, we aimed to further refine the LRCN architecture to improve its performance on diverse video datasets.

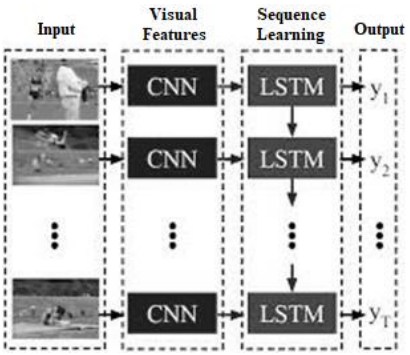

Figure 1. The LRCN architectures

3.3. Our model

Our proposed model builds upon the foundational principles of LRCN, but introduces several key enhancements to improve performance in video classification tasks. While LRCN combines CNNs and LSTM networks [23], we focus on optimizing certain architectural elements to boost the model's ability to handle complex video datasets.

In our model, we have reconfigured the size of the convolutional filters. The standard LRCN architecture typically uses fixed filter sizes, but we explore the impact of varying these filter sizes to better capture spatial features at different granularities. Larger filters allow the network to capture broader spatial patterns, while smaller filters help in focusing on finer details. This dynamic adjustment enables our model to better adapt to the complexities of different video datasets. Similar to LRCN, our model utilizes time-distributed convolutional layers. These layers operate independently on each frame of the video sequence, ensuring that the spatial features are extracted frame by frame without disturbing the temporal order. By maintaining the sequence's integrity, we preserve the temporal dynamics between frames, which is crucial for understanding motion and actions in videos. Figure 2 shows the overview of the proposed video classification pipeline.

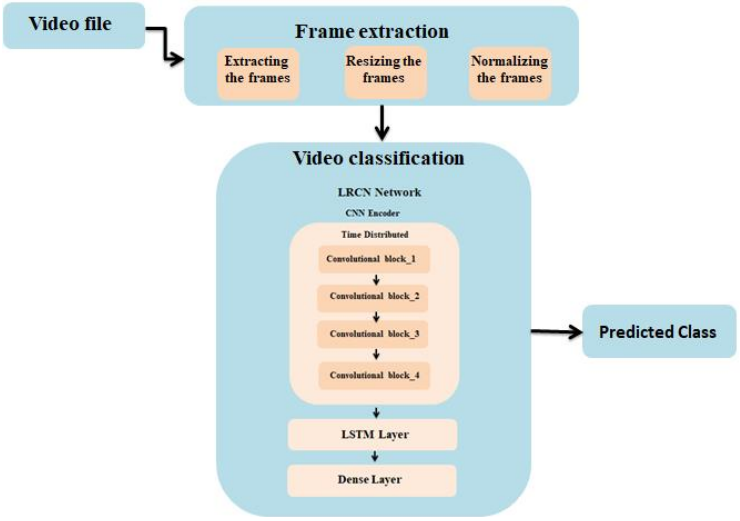

Figure 2. Overview of the proposed video classification pipeline

As illustrated in Figure 2, the proposed video classification framework follows a structured pipeline that integrates video preprocessing and deep spatio-temporal learning. The process starts with a raw video

file, from which a fixed number of frames are extracted to preserve temporal consistency across all samples. These frames are then subjected to preprocessing steps, including frame extraction, resizing, and normalization, ensuring standardized input dimensions and stable training behavior.

After preprocessing, the processed frames are forwarded to the LRCN-based video classification module. Each frame is independently processed using time-distributed convolutional layers, allowing spatial feature extraction while maintaining the temporal order of the video sequence. The CNN encoder is composed of multiple Conv2D layers with progressively increasing numbers of filters (e.g., 16, 32, 64, and 128), each followed by MaxPooling2D layers to reduce spatial resolution and computational complexity. This hierarchical convolutional design enables the network to capture low-level features such as edges and textures in early layers, while deeper layers focus on more abstract and discriminative spatial representations.

The extracted frame-level features are then flattened and arranged into a temporal feature sequence, which is fed into an LSTM layer to model long-term temporal dependencies and motion dynamics across frames. Finally, a dense layer followed by a classification layer produces the predicted class label for the input video. This pipeline effectively combines spatial and temporal learning, providing a robust framework for video classification. Figure 3 details the internal architecture of the proposed LRCN network, including the convolutional blocks and the temporal modeling components.

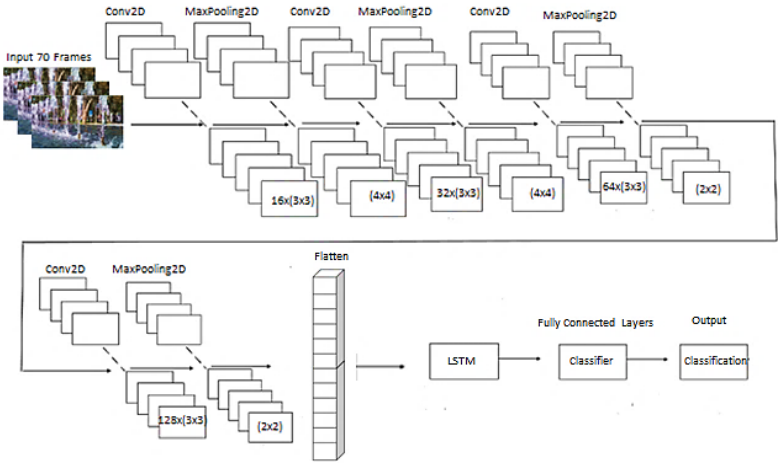

Figure 3. Detailed architecture of the proposed LRCN model

4. RESULTS AND DISCUSSION

The results demonstrate the effectiveness of the proposed model, which outperforms both ConvLSTM and LRCN across all datasets DynTex, UCF11, and UCF50 achieving superior accuracy levels. This underscores the robustness and adaptability of the proposed model in addressing diverse video classification challenges. Table 1 shows the comparison of video classification methods.

Table 1. Comparison of video classification methods

Method	DynTex	UCF11	UCF50
ConvLSTM [23]	0.56	0.62	0.79
LRCN [7]	0.71	0.77	0.93
BT-LSTM [24]	N/A	0.85	N/A
DEEPEYE [25]	N/A	0.86	N/A
TR-LSTM [26]	N/A	0.87	N/A
KCP-LSTM [27]	N/A	0.88	0.87
HT-LSTM [28]	N/A	N/A	0.76
Fusion feature [29]	N/A	N/A	0.91
Proposed model	0.90	0.92	0.94

The table provides a comprehensive comparison of multiple deep learning models evaluated on three widely used video classification datasets: DynTex, UCF11, and UCF50. The models compared include well-established baseline architectures such as ConvLSTM and LRCN, as well as more advanced and specialized models, including BT-LSTM, DEEPEYE, TR-LSTM, KCP-LSTM, HT-LSTM, and fusion

feature-based methods. Performance metrics are presented in terms of classification accuracy. On the DynTex dataset, only two models ConvLSTM and LRCN report results, with classification accuracies of 0.56 and 0.71, respectively. These results demonstrate the limitations of earlier recurrent architectures in capturing the subtle temporal dynamics of texture-based sequences. In contrast, the proposed model significantly outperforms these methods with an accuracy of 0.90, suggesting its superior ability to model temporal dependencies and spatial patterns inherent in dynamic textures. The UCF11 dataset sees broader model coverage. Traditional methods like ConvLSTM and LRCN achieve accuracies of 0.62 and 0.77, respectively, while more recent models such as BT-LSTM (0.85), DEEPEYE (0.86), TR-LSTM (0.87), and KCP-LSTM (0.88) demonstrate incremental improvements by integrating deeper temporal modeling and attention mechanisms. Nevertheless, the proposed model achieves the highest accuracy of 0.92, underscoring its enhanced generalization and feature representation capabilities in the context of real-world human action videos. On the more challenging UCF50 dataset, which includes a larger and more diverse set of action classes, performance results continue to improve with more advanced architectures. LRCN achieves a strong accuracy of 0.93, while KCP-LSTM and the fusion feature methods reach 0.87 and 0.91, respectively. Notably, the proposed model leads again with the best overall performance of 0.94, indicating its robustness and adaptability across different types of video data, even in the presence of intra-class variability and complex motion patterns. Overall, the proposed model consistently outperforms all compared methods across the three datasets. Its superior performance can be attributed to enhanced temporal modeling, better spatial-temporal feature fusion, and possibly the use of more effective training strategies or architectural innovations. These results validate the model's effectiveness and make it a strong candidate for further applications in video understanding tasks such as surveillance, activity recognition, and content-based video retrieval.

As illustrated in Figure 4, the comparison demonstrates that the proposed architecture consistently outperforms the baseline LRCN across all evaluated datasets. On the DynTex dataset, the proposed architecture achieves a substantial improvement in accuracy, highlighting its enhanced ability to capture complex dynamic texture patterns. Similar significant gains are observed on the UCF11 dataset, reflecting more effective temporal modeling of human actions. On the UCF50 dataset, although the improvement is smaller, the proposed architecture still attains the highest accuracy, confirming its robustness on larger and more diverse action categories. Overall, these results validate the effectiveness of the proposed architectural enhancements and demonstrate their contribution to improved spatio-temporal feature representation in video classification tasks.

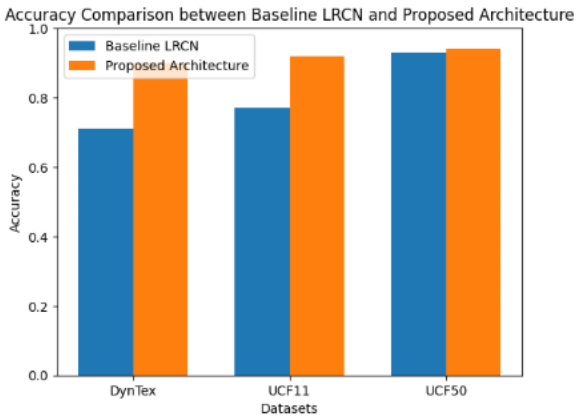

Figure 4. Comparison between baseline LRCN and new proposed architecture

5. CONCLUSION AND FUTURE WORK

In this study, we proposed an enhanced LRCN for video classification. Through targeted architectural refinements and improved temporal modeling, our model effectively captured detailed spatio-temporal patterns. It achieved accuracies of 0.90 on DynTex (+26.8% over the standard LRCN), 0.92 on UCF11 (+19.5%), and 0.94 on UCF50 (+1.1%), consistently outperforming ConvLSTM, LRCN, and other state-of-the-art methods. These results demonstrate the robustness and versatility of the enhanced LRCN in capturing complex spatial and temporal dynamics across diverse video datasets.

Looking ahead, future research could focus on integrating self-attention mechanisms or transformer blocks into the LRCN architecture to further enhance temporal feature modeling. This integration would allow the model to capture long-range dependencies more effectively across video frames, potentially

improving action recognition in complex or long-duration sequences. Additionally, exploring cross-domain transfer learning could evaluate the model's generalization ability to unseen datasets and diverse video types, directly addressing current limitations of the enhanced LRCN. Overall, these directions provide promising avenues to advance the development of more accurate, adaptable, and efficient video classification models.

FUNDING INFORMATION

The authors declare that this research was conducted without any research grant or contract.

ACKNOWLEDGMENTS

The authors would like to thank all those who contributed, directly or indirectly, to the completion of this work. We also acknowledge the constructive comments and suggestions provided by the reviewers, which helped improve the quality of this manuscript.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su
Benzyane manal	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Azrour mourade										✓		✓
Agoujil said										✓		✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The DynTex dataset is not publicly available but can be accessed upon reasonable request from the corresponding author. The UCF11 and UCF50 datasets are publicly accessible from their official repositories.

REFERENCES

[1] M. Benzyane, M. Azrour, I. Zeroual, and S. Agoujil, "State-of-the-art methods for dynamic texture classification: a comprehensive review," in *World Sustainability Series*, vol. Part F2570, 2024, pp. 1–13.
[2] L. Arrotta, G. Civitarese, X. Chen, J. Cumin, and C. Bettini, "Multi-subject human activities: a survey of recognition and evaluation methods based on a formal framework," *Expert Systems with Applications*, vol. 267, p. 126178, Apr. 2025, doi: 10.1016/j.eswa.2024.126178.
[3] D. Kong, S. Hu, and G. Zhao, "MV-STCNet: Breast cancer diagnosis using spatial and temporal dual-attention guided classification network based on multi-view ultrasound videos," *Biomedical Signal Processing and Control*, vol. 87, p. 105541, Jan. 2024, doi: 10.1016/j.bspc.2023.105541.
[4] Y. Lu *et al.*, "Video surveillance-based multi-task learning with swin transformer for earthwork activity classification," *Engineering Applications of Artificial Intelligence*, vol. 131, p. 107814, May 2024, doi: 10.1016/j.engappai.2023.107814.
[5] M. A. Uddin *et al.*, "Deep learning-based human activity recognition using CNN, ConvLSTM, and LRCN," *International Journal of Cognitive Computing in Engineering*, vol. 5, pp. 259–268, 2024, doi: 10.1016/j.ijcce.2024.06.004.
[6] B. Koger, A. Deshpande, J. T. Kerby, J. M. Graving, B. R. Costelloe, and I. D. Couzin, "Quantifying the movement, behaviour and environmental context of group-living animals using drones and computer vision," *Journal of Animal Ecology*, vol. 92, no. 7, pp. 1357–1371, Jul. 2023, doi: 10.1111/1365-2656.13904.
[7] M. Benzyane, M. Azrour, I. Zeroual, and S. Agoujil, "Investigating the influence of convolutional operations on LSTM networks in video classification," *Data and Metadata*, vol. 2, p. 152, Dec. 2023, doi: 10.56294/dm2023152.
[8] Z. Ruiye, "Volleyball training video classification description using the BiLSTM fusion attention mechanism," *Heliyon*, vol. 10, no. 15, p. e34735, Aug. 2024, doi: 10.1016/j.heliyon.2024.e34735.

[9] A. Raza *et al.*, "An improved deep convolutional neural network-based YouTube video classification using textual features," *Heliyon*, vol. 10, no. 16, p. e35812, Aug. 2024, doi: 10.1016/j.heliyon.2024.e35812.
[10] X. Li, H. Wu, M. Li, and H. Liu, "Multi-label video classification via coupling attentional multiple instance learning with label relation graph," *Pattern Recognition Letters*, vol. 156, pp. 53–59, Apr. 2022, doi: 10.1016/j.patrec.2022.01.003.
[11] Y. Qiao, Y. Guo, K. Yu, and D. He, "C3D-ConvLSTM based cow behaviour classification using video data for precision livestock farming," *Computers and Electronics in Agriculture*, vol. 193, p. 106650, Feb. 2022, doi: 10.1016/j.compag.2021.106650.
[12] L. Zhang *et al.*, "Video Deepfake classification using particle swarm optimization-based evolving ensemble models," *Knowledge-Based Systems*, vol. 289, p. 111461, Apr. 2024, doi: 10.1016/j.knosys.2024.111461.
[13] C. L. Angelina *et al.*, "Classification of pancreatic cystic lesions using resnet deep learning network in confocal laser endomicroscopy videos," *Procedia Computer Science*, vol. 234, pp. 357–363, 2024, doi: 10.1016/j.procs.2024.03.015.
[14] A. Ouessai and M. Keche, "IMM/EKF filter based classification of real-time freeway video traffic without learning," *Transportation Letters*, vol. 14, no. 6, pp. 610–621, Jul. 2022, doi: 10.1080/19427867.2021.1913304.
[15] M. Benzyane, I. Zeroual, M. Azrour, and S. Agoujil, "Convolutional long short-term memory network model for dynamic texture classification: a case study," in *Lecture Notes in Networks and Systems*, vol. 637 LNNS, 2023, pp. 383–395.
[16] C. Zhao, J. G. Han, and X. Xu, "CNN and RNN based neural networks for action recognition," *Journal of Physics: Conference Series*, vol. 1087, no. 6, p. 062013, Sep. 2018, doi: 10.1088/1742-6596/1087/6/062013.
[17] C. Wu, Y. Sang, and Y. Gao, "Extreme learning machine combining hidden-layer feature weighting and batch training for classification," *Neural Processing Letters*, vol. 55, no. 8, pp. 10951–10973, Dec. 2023, doi: 10.1007/s11063-023-11358-2.
[18] R. Vrskova, P. Kamencay, R. Hudec, and P. Sykora, "A new deep-learning method for human activity recognition," *Sensors*, vol. 23, no. 5, p. 2816, Mar. 2023, doi: 10.3390/s23052816.
[19] J. Choi, J. S. Lee, M. Ryu, G. Hwang, G. Hwang, and 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 2022 - Conference Proceedings*, Jun. 2022, pp. 1–6, doi: 10.1109/MeMeA54994.2022.9856417.
[20] S. D. Khan, G. Vizzari, and S. Bandini, "Identifying sources and sinks and detecting dominant motion patterns in crowds," *Transportation Research Procedia*, vol. 2, pp. 195–200, 2014, doi: 10.1016/j.trpro.2014.09.030.
[21] J. Di and H. Liu, "Research of moving target tracking technology based on LRCN," in *2017 International Conference on Computer Systems, Electronics and Control, ICCSEC 2017*, Dec. 2018, pp. 789–792, doi: 10.1109/ICCSEC.2017.8446988.
[22] V. B. Vinusha, V. Indhuja, M. V. Reddy, N. Nikhitha, and P. Pramila, "Suspicious activity detection using LRCN," in *Proceedings - 5th International Conference on Smart Systems and Inventive Technology, ICSSIT 2023*, Jan. 2023, pp. 1463–1470, doi: 10.1109/ICSSIT55814.2023.10061045.
[23] M. Benzyane, M. Azrour, I. Zeroual, and S. Agoujil, "Exploring the Impact of Convolutions on LSTM networks for video classification," in *Lecture Notes in Networks and Systems*, vol. 838 LNNS, 2024, pp. 21–26.
[24] J. Ye *et al.*, "Learning compact recurrent neural networks with block-term tensor decomposition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 9378–9387, doi: 10.1109/CVPR.2018.00977.
[25] Y. Cheng, G. Li, H.-B. Chen, S. X.-D. Tan, and H. Yu, "DEEPEYE: a compact and accurate video comprehension at terminal devices compressed with quantization and tensorization," *arXiv: arXiv:1805.07935*, 2018, doi: 10.48550/arXiv.1805.07935.
[26] Y. Pan *et al.*, "Compressing recurrent neural networks with tensor ring for action recognition," *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, vol. 33, no. 01, pp. 4683–4690, Jul. 2019, doi: 10.1609/aaai.v33i01.33014683.
[27] D. Wang *et al.*, "Kronecker CP decomposition with fast multiplication for compressing RNNs," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 5, pp. 2205–2219, May 2023, doi: 10.1109/TNNLS.2021.3105961.
[28] B. Wu, D. Wang, G. Zhao, L. Deng, and G. Li, "Hybrid tensor decomposition in neural network compression," *Neural Networks*, vol. 132, pp. 309–320, Dec. 2020, doi: 10.1016/j.neunet.2020.09.006.
[29] D. Wang, J. Yang, and Y. Zhou, "Human action recognition based on multi-mode spatial-temporal feature fusion," in *FUSION 2019 - 22nd International Conference on Information Fusion*, Jul. 2019, pp. 1–7, doi: 10.23919/fusion43075.2019.9011361.

BIOGRAPHIES OF AUTHORS

Manal Benzyane is currently a fourth-year Ph.D. student specializing in "Spatial Descriptors for Dynamic Texture Recognition". She earned her Master's degree (M.Sc.) in 2021 in Systems for Information and Decision Imagery (SIDI). Her research focuses on video classification, video recognition, and spatiotemporal deep learning models. She has authored several scientific publications, including journal articles and book chapters, such as: "Investigating the Influence of Convolutional Operations on LSTM Networks in Video Classification", "Dynamic Texture Classification Using ConvLSTM and Video Optical Flow Analysis", "Convolutional Long Short-Term Memory Network Model for Dynamic Texture Classification: A Case Study", and "Exploring the Impact of Convolutions on LSTM Networks for Video Classification". Her academic interests include dynamic texture analysis, deep neural networks, LSTM-based models, and spatiotemporal feature extraction. She is actively involved in the scientific community through contributions to international conferences and peer-reviewed publications. Learning, text classification, natural language processing, and machine learning. She can be contacted at email: m.benzyane@edu.umi.ac.ma.

Prof. Dr. Mourade Azrour received his Ph.D. from Faculty of sciences and Techniques, Moulay Ismail University of Meknes, Morocco. He has received his MS in computer and distributed systems from Faculty of Sciences, Ibn Zouhr University, Agadir, Morocco in 2014. Mourade currently works as computer sciences professor at the Department of Computer Science, Faculty of Sciences and Techniques, Moulay Ismail University of Meknès. His research interests include authentication protocol, computer security, internet of things, smart systems, machine learning and so ones. Mourade is member of the member of the scientific committee of numerous international conferences. He is also a reviewer of various scientific journals. He has published more than 137 scientific papers and book chapters. Mourade has edited many scientific books for example: "IoT, Machine Learning and Data Analytics for Smart Healthcare", "Blockchain and Machine Learning for IoT Security", "IoT and Smart Devices for Sustainable Environment", "Advanced Technology for Smart Environment and Energy", and so ones. Finally, he has served as guest editor in journals "EAI Endorsed Transactions on Internet of Things", "Tsinghua Science and Technology", "Applied Sciences MDPI" and "Sustainability MDPI". He can be contacted at email: mo.azrour@umi.ac.ma.

Prof. Said Agoujil received his Ph.D. from Faculty of sciences and Techniques Marrakech, Cadi Ayyad University of Marrakech, Morocco. Said Agoujil currently works as computer sciences professor at the Department of Computer Science, National Business School of Commerce and Management, Moulay Ismail University of Meknes. His research interests include Numerical Analysis, Mobile network, Internet of things, Smart systems, Machine learning and so ones. Said is member of the member of the scientific committee of numerous international conferences. He is also a reviewer of various scientific journals. He has published more than 37 scientific papers and book chapters. He can be contacted at email: agoujil@gmail.com.