Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 01-16
DOI: https://doi.org/10.51903/jtie.v5i1.441

Efficient Temporal Segmentation and Classification of Short-Form Video Content Using a Lightweight CNN-LSTM Architecture

Ben Liu Tan[1], Chstina Angel Liem*[2], Mohamed Amen[3]
Email: chstliem@gmail.com
[1,2,3]*Monash University, Melbourne, Victoria, Australia, 3800*
*Corresponding Author

Abstract
The exponential rise of short-form video platforms such as TikTok, Instagram Reels, and YouTube Shorts has transformed digital content consumption patterns, creating both opportunities and challenges in media analysis. One critical need is the efficient segmentation and classification of temporal segments within these videos to enable applications in content moderation, targeted advertising, and audience behavior research. This study proposes a lightweight deep learning architecture that integrates Convolutional Neural Networks (CNN) for visual feature extraction and Long Short-Term Memory (LSTM) networks for temporal sequence modeling. The proposed CNN-LSTM framework is optimized for computational efficiency while maintaining high classification accuracy, making it suitable for deployment in resource-constrained environments. Experimental evaluations on a curated short-form video dataset show that the model achieves competitive performance compared with larger architectures while significantly reducing memory usage and inference time. Furthermore, the temporal segmentation module effectively isolates meaningful visual-audio segments, enabling more precise classification outcomes. The results highlight the potential of lightweight architectures to address the scalability demands of modern video analysis systems without sacrificing accuracy. This research contributes to the growing discourse on efficient multimedia processing by bridging the gap between high-performance models and practical, real-time applications in the evolving short-form video ecosystem.

Keywords: *Lightweight deep learning, Temporal segmentation, Short-form video classification, CNN-LSTM, Multimedia content analysis.*

I. INTRODUCTION

The rapid proliferation of short-form video platforms such as TikTok, Instagram Reels, and YouTube Shorts has transformed the dynamics of digital media consumption, driving new paradigms for audience engagement and content delivery (Zhang et al., 2022). These platforms thrive on delivering highly engaging, bite-sized audiovisual content that caters to decreasing attention spans and the demand for instant gratification (Raharjo et al., 2024). As user-generated videos become a dominant mode of online communication, the challenge of automatically segmenting and classifying their temporal structures becomes increasingly significant for applications such as content moderation, personalized recommendation, and targeted advertising (Dzhoha et al., 2025; Montefalcon et al., 2021). Unlike long-form videos, short-form content presents unique challenges due to rapid scene changes, multimodal cues, and a condensed narrative structure, requiring efficient processing pipelines that operate under strict computational constraints. Addressing these challenges requires integrating advanced deep learning methods with optimized architectures that balance performance and resource efficiency (Liu et al., 2024; Mittal, 2024).

Received on September, 2025; Revised on November, 2025; Accepted on December, 2025; Published on April, 2026

Temporal segmentation, the process of dividing a video into semantically coherent segments, plays a pivotal role in understanding video narratives and enabling downstream tasks such as classification or summarization (Pasquarella et al., 2022). Modern research in temporal action segmentation has focused extensively on methods for long and untrimmed videos, often leveraging heavy architectures that are computationally expensive (Ding et al., 2023; Singhania et al., 2023). However, short-form videos require approaches that can rapidly and accurately detect scene or action boundaries without introducing significant latency. Lightweight deep learning models have emerged as a promising solution for resource-constrained environments, offering the potential for real-time processing on edge devices while maintaining high accuracy (Wei et al., 2020; Zhao et al., 2020). This is particularly relevant in the context of massive-scale short-form content streams, where scalability and energy efficiency are critical for practical deployment.

The fusion of Convolutional Neural Networks (CNN) for spatial feature extraction with Long Short-Term Memory (LSTM) networks for temporal modeling has shown strong potential in domains such as stock price forecasting (Lu et al., 2020), environmental modeling (Elmaz et al., 2021), and event prediction (Rostamian & O'Hara, 2022). In the video domain, CNN-LSTM hybrids have demonstrated the ability to capture fine-grained spatial details while learning sequential dependencies across frames, making them suitable for both segmentation and classification tasks (Ullah et al., 2024). Yet, most existing implementations prioritize accuracy over efficiency, making them unsuitable for real-time short-form video analysis at scale. This gap underscores the need for novel lightweight CNN-LSTM designs tailored specifically for short-form content, where both computational constraints and multimodal integration (e.g., audio-visual features) must be considered simultaneously.

The significance of efficient temporal segmentation is further amplified by the rise of multimodal content understanding, where visual, audio, and textual cues interact in complex ways to convey meaning (Athar et al., 2023; Grammatikopoulou et al., 2023). Short-form videos often rely heavily on synchronized audio effects, quick visual transitions, and text overlays to enhance narrative impact, which makes purely visual analysis insufficient. Incorporating temporal reasoning and multimodal embeddings can substantially improve segmentation and classification performance (Chen et al., 2020; Huang et al., 2020). Furthermore, emerging research highlights the importance of addressing dataset biases, cold-start scenarios, and fairness in short-form content recommendation systems, reinforcing the need for segmentation models that are both efficient and robust across diverse content types (Dzhoha et al., 2025).

Lightweight architectures have seen success across various domains beyond video processing, including cardiovascular disease detection (Shuvo et al., 2021), intrusion detection (Doriguzzi-Corin et al., 2020), and edge-based surveillance (Wang et al., 2020; Zhao et al., 2020). These solutions demonstrate that well-optimized deep learning pipelines can achieve near state-of-the-art accuracy while drastically reducing memory footprint and inference latency. In the context of short-form video analysis, this capability enables deployment in mobile applications, embedded devices, and large-scale streaming platforms where hardware resources are limited. Recent surveys emphasize that the design of lightweight models must consider not only architectural compression but also training strategies, data efficiency, and hardware-specific optimizations to achieve optimal results (Liu et al., 2024; Mittal, 2024).

From a broader perspective, the development of efficient segmentation and classification models also aligns with ethical and societal concerns surrounding short-form video ecosystems. Privacy issues, misinformation, and harmful content have been widely documented as challenges for platform governance (Hariguna et al., 2022). Automated systems that rapidly analyze and categorize short-form content can be a critical component of proactive moderation frameworks, helping identify policy violations before they reach large audiences. Furthermore, content segmentation can support educational and cultural applications, such as adaptive learning systems, by allowing educators to curate micro-content based on specific learning objectives (Bahroun et al., 2023). Thus, research into lightweight CNN-LSTM architectures is not only a technical endeavor but also a socially relevant one.

In light of these developments, this study introduces an efficient CNN-LSTM architecture for temporal segmentation and classification of short-form videos. The proposed model integrates spatial and temporal feature learning with computational efficiency, leveraging a design optimized for both speed and accuracy. It addresses the lack of research focusing specifically on short-form content, where traditional segmentation approaches struggle to maintain responsiveness under large-scale, real-time demands. Through experiments on curated datasets of TikTok, Reels, and Shorts videos, the study demonstrates that the proposed architecture achieves competitive accuracy while significantly reducing model size and inference time. By bridging the gap between high-performance video analysis and resource-constrained deployment scenarios, this research contributes to the growing discourse on lightweight deep learning in multimedia content processing.

II. LITERATURE REVIEW

The analysis of short-form video content has emerged as a rapidly expanding research domain, driven by the growth of platforms such as TikTok, Instagram Reels, and YouTube Shorts (Zhang

et al., 2022). Unlike traditional long-form content, short-form videos often feature condensed narratives, rapid transitions, and heavy reliance on multimodal cues such as music, visual overlays, and captions. These characteristics require novel approaches to temporal segmentation and classification that are both accurate and computationally efficient. Conventional video analysis pipelines often depend on large, resource-intensive models that may not be practical for real-time, large-scale deployment (Liu et al., 2024; Mittal, 2024). As a result, recent studies have explored lightweight deep learning architectures capable of operating in resource-constrained environments while preserving high accuracy (Wei et al., 2020; Zhao et al., 2020).

Temporal action segmentation forms the foundation for understanding complex video sequences, enabling downstream tasks such as activity recognition, event detection, and content classification (Ding et al., 2023). Early methods for segmentation were predominantly rule-based or relied on handcrafted features, but advances in deep learning have enabled the modeling of temporal dependencies using architectures such as Temporal Convolutional Networks (TCN) and recurrent neural networks (RNN) (Li et al., 2021; Singhania et al., 2023). Recent research has expanded to address domain-specific challenges, such as handling noisy annotations or performing segmentation under weak supervision (Chen et al., 2020). Furthermore, multimodal temporal segmentation approaches, which leverage both visual and audio streams, have shown potential in improving accuracy and robustness (Athar et al., 2023; Huang et al., 2020). However, these techniques often incur high computational costs, underscoring the need for efficient solutions.

Lightweight deep learning models have demonstrated remarkable adaptability across a variety of domains, including intrusion detection (Doriguzzi-Corin et al., 2020), medical diagnostics (Shuvo et al., 2021; Wei et al., 2020), and communications (Wang et al., 2020). In the video domain, lightweight CNN-based models have been optimized for edge devices, enabling real-time inference without dependence on high-performance servers (Liu et al., 2024; Mittal, 2024). These architectures commonly incorporate techniques such as network pruning, quantization, and knowledge distillation to reduce complexity while retaining predictive power. The integration of lightweight CNN with sequence modeling architectures such as LSTM has shown strong results in tasks like load forecasting (Ullah et al., 2024), stock price prediction (Lu et al., 2020), and event prediction in dynamic environments (Rostamian & O'Hara, 2022). The same principles apply to short-form video segmentation, where low-latency performance is critical.

The fusion of CNN for spatial feature extraction and LSTM for temporal modeling has proven effective in learning both short-term and long-term dependencies within sequential data (Elmaz et al., 2021). In video applications, CNN-LSTM hybrids can simultaneously process frame-level spatial patterns and capture temporal dynamics across scenes or actions. This dual capability

makes them particularly suitable for applications that require both segmentation and classification (Grammatikopoulou et al., 2023). Despite these strengths, most CNN-LSTM models in the literature are designed for high-performance computing environments and are rarely optimized for lightweight, real-time scenarios (Liu et al., 2024; Mittal, 2024). Consequently, there remains an underexplored research gap in adapting CNN-LSTM architectures to efficiently process short-form video content.

Beyond the technical dimension, understanding short-form videos also involves examining their sociotechnical implications. Studies have highlighted how these platforms influence user behavior, privacy concerns, and content dynamics (Hariguna et al., 2022; Montefalcon et al., 2021). Content analysis techniques, which involve systematic categorization and interpretation of multimedia data, have been widely applied to uncover patterns in user-generated content across contexts such as health communication (O'Hagan et al., 2021; Rufai & Bunce, 2020), mental health (Oyetunji et al., 2021), and emerging virtual environments (Narin, 2021). While these approaches traditionally rely on manual coding or statistical models, the integration of automated segmentation and classification methods offers the potential to scale such analyses to millions of videos, enabling richer and more timely insights (Kleinheksel et al., 2020).

The problem of recommendation bias and cold-start scenarios in short-form video platforms further emphasizes the importance of accurate and efficient video understanding models (Dzhoha et al., 2025). Recommendation systems increasingly rely on multimodal embeddings that capture visual, audio, and textual features for improved personalization. Temporal segmentation enhances these systems by providing more granular representations of user-consumed content, thereby improving the relevance of recommendations. Techniques such as spatio-temporal embeddings have been proposed to enhance instance segmentation in videos, enabling better tracking and classification of dynamic objects (Athar et al., 2023). However, applying such advanced embeddings to lightweight architectures for short-form video remains an open research area.

Research on temporal reasoning in videos has explored multiple strategies to improve segmentation accuracy. Graph-based temporal reasoning models have been proposed to capture complex dependencies between actions across time (Huang et al., 2020), while timestamp-supervised learning approaches aim to reduce annotation costs (Li et al., 2021). Domain adaptation techniques, such as self-supervised temporal domain adaptation, have also been shown to enhance model generalization across diverse datasets (Chen et al., 2020). While these advancements have significantly improved segmentation performance, their high computational demands make direct application to short-form content challenging. Adapting these methods into lightweight, efficient architectures could provide a promising direction for future research.

In parallel, lightweight deep learning for edge-based surveillance and multimedia analytics has gained traction in recent years, demonstrating that efficiency does not necessarily come at the cost of accuracy (Wei et al., 2020; Zhao et al., 2020). Studies such as LightAMC (Wang et al., 2020) and CardioXNet (Shuvo et al., 2021) Exemplify how targeted model design can deliver high-performance results under strict resource constraints. In the context of short-form videos, such architectural principles can be adapted to ensure fast, accurate segmentation and classification without requiring cloud-based computation. This shift toward on-device processing also aligns with privacy-preserving computing trends, which are increasingly important in regulating user-generated content.

Taken together, the literature highlights three critical insights for this research. First, there is a clear demand for temporal segmentation methods tailored to the unique dynamics of short-form videos, incorporating multimodal cues and rapid scene changes. Second, CNN-LSTM architectures offer a promising framework for joint segmentation and classification but require substantial optimization for resource-constrained deployment. Third, when applied thoughtfully, lightweight deep learning principles can enable scalable, real-time video analysis while maintaining accuracy and preserving privacy. These insights collectively form the foundation for the present study, which seeks to develop and evaluate an efficient CNN-LSTM architecture specifically designed for temporal segmentation and classification of short-form video content.

III. RESEARCH METHOD

This study adopts a systematic research methodology designed to develop and evaluate a lightweight CNN-LSTM architecture for temporal segmentation and classification of short-form videos. Building upon the challenges and opportunities identified in the literature review, the methodology integrates principles from lightweight deep learning (Liu et al., 2024; Mittal, 2024) and temporal segmentation research (Ding et al., 2023; Singhania et al., 2023) into a coherent experimental framework. The approach emphasizes efficiency, scalability, and accuracy while considering the unique multimodal characteristics of short-form content (Zhang et al., 2022). Figure 1 presents the research flow, illustrating the sequence of stages from data collection to model evaluation.

The research process begins with dataset preparation, which involves gathering publicly available short-form videos containing varied content types, transitions, and audio-visual cues. Following ethical data collection guidelines and ensuring privacy compliance (Hariguna et al., 2022), the videos are preprocessed through frame extraction, audio separation, and annotation. Temporal annotations are created to mark the start and end of distinct actions or segments, following timestamp-supervised methods suggested by (Li et al., 2021). This structured preparation ensures

that the dataset is both representative of real-world scenarios and compatible with the intended segmentation and classification tasks.

A. Dataset and Experimental Setup

This study uses a curated short-form video dataset collected from three major social media platforms: TikTok, Instagram Reels, and YouTube Shorts. The final dataset consists of 2,400 short-form videos. The average video duration is 18.6 seconds, ranging from 5 to 45 seconds, ensuring sufficient temporal variability for segmentation and classification tasks. The dataset is categorized into six content classes: entertainment, lifestyle, education, sports, news, and advertising. The class distribution was maintained as balanced as possible to reduce classification bias (Matuan et al., 2026). Each video was manually annotated using a timestamp-based labeling scheme, in which the start and endpoints of each semantic action segment were marked at the frame level.

The annotation process involved three independent annotators with prior experience in multimedia data labeling. To ensure labeling reliability and consistency, inter-annotator agreement was evaluated using Cohen's Kappa coefficient, yielding a score of 0.86, indicating a high level of consistency. Any disagreements among annotators were resolved by majority voting. For model training and evaluation, the dataset was divided into three subsets in an 80:10:10 ratio: 1,920 videos for training, 240 for validation, and 240 for testing. Stratified sampling was applied to preserve the proportional distribution of classes across all subsets.

All video frames were resized to a fixed spatial resolution of 224 × 224 pixels and uniformly sampled at 30 frames per second (FPS) to ensure computational stability and a fair comparison of models. All experiments were conducted on a workstation equipped with an Intel Core i9-12900K CPU, 32 GB RAM, and an NVIDIA RTX 3080 GPU. Inference performance on this setup was measured, showing an average processing time of 0.38 seconds per video. For edge deployment evaluation, the model was tested on a Jetson Nano device, achieving real-time processing speed of 78 FPS. The proposed model was implemented using Python and TensorFlow. These specifications ensure that the model's efficiency, real-time inference capability, and feasibility for edge-device deployment are clearly documented.

B. Data Preprocessing and Feature Extraction

The preprocessing stage incorporates spatial and temporal normalization steps to standardize the visual and auditory inputs. Frames are resized to a fixed resolution to reduce computational overhead, consistent with lightweight processing principles (Wei et al., 2020; Zhao et al., 2020). Audio features are extracted in the form of Mel-frequency cepstral coefficients (MFCC) to

complement visual data, enabling multimodal feature fusion (Athar et al., 2023; Huang et al., 2020). Data augmentation techniques such as random cropping, horizontal flipping, and color jittering are applied to improve generalization while avoiding overfitting, in line with best practices in video deep learning.

C. Model Architecture

The core model architecture integrates a lightweight convolutional neural network (CNN) for spatial feature extraction and a long short-term memory (LSTM) network for temporal modeling. The CNN component is based on a pruned and quantized variant of a standard architecture to ensure low latency on resource-constrained devices (Liu et al., 2024; Mittal, 2024). The extracted spatial features are passed to an LSTM layer, which captures both short-term and long-term dependencies in sequential video frames, following the hybrid design principles outlined by (Elmaz et al., 2021; Rostamian & O'Hara, 2022). This combination is chosen because it allows the model to learn frame-level patterns and temporal transitions simultaneously, addressing the unique segmentation needs of short-form videos.

D. Temporal Segmentation Strategy

Temporal segmentation within the model is achieved using a sequence-labeling approach, in which each frame is assigned an action class. The output layer produces probabilities for each temporal class, enabling the identification of segment boundaries. To enhance segmentation accuracy, the architecture incorporates spatio-temporal embeddings inspired by the STEm-Seg approach (Athar et al., 2023) and temporal reasoning mechanisms from graph-based methods (Huang et al., 2020). Prior findings support the decision to incorporate these techniques, which improve robustness to abrupt scene changes, a common characteristic of short-form videos.

E. Training Strategy

Model training is conducted using a supervised learning paradigm with cross-entropy loss for classification and a boundary-sensitive loss function for segmentation accuracy. The training process employs the Adam optimizer with an adaptive learning rate schedule to balance convergence speed and stability. To ensure fairness and prevent bias in classification, especially in user-generated content, the training set is balanced across categories and scenarios, following recommendations from (Dzhoha et al., 2025). The implementation is carried out in a GPU-enabled environment for initial training, followed by inference benchmarking on edge devices to assess the feasibility of real-world deployment.

F. Performance Evaluation

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 01-16
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Performance evaluation involves both quantitative and qualitative assessments. Quantitative metrics include accuracy, precision, recall, and F1-score for classification, along with mean Intersection-over-Union (mIoU) and edit score for temporal segmentation (Pasquarella et al., 2022). Qualitative evaluation involves visually inspecting segmentation timelines to verify that predicted segment boundaries align with human-annotated ground truth. Table 1 summarizes the evaluation metrics, their definitions, and their relevance to the present study. The combination of these metrics ensures a holistic understanding of the model's performance.

Table 1. Evaluation Metrics Used for Temporal Segmentation and Classification

Metric	Definition	Relevance to Study
Accuracy	Ratio of correctly classified frames to total frames	Measures overall classification correctness
Precision	True positives divided by total predicted positives	Evaluates the correctness of positive predictions
Recall	True positives divided by total actual positives	Evaluates the completeness of positive predictions
F1-score	Harmonic mean of precision and recall	Balances precision and recall performance
mIoU	Average IoU across all classes	Assesses segmentation quality across categories
Edit Score	Normalized Levenshtein distance between predicted and ground truth sequences	Measures temporal alignment quality

G. Deployment and Edge Testing

The final stage of the methodology is model deployment testing. This stage involves deploying the trained model on low-power hardware, such as a Raspberry Pi or a Jetson Nano, to simulate real-world operating conditions. The primary objective is to evaluate inference speed, memory usage, and real-time segmentation feasibility without reliance on cloud processing, in line with privacy-preserving and edge-computing principles (Doriguzzi-Corin et al., 2020; Shuvo et al., 2021). Insights from this stage directly inform recommendations for scaling and integrating the model into production systems. By integrating these components, the methodology ensures that the proposed CNN–LSTM framework is not only accurate but also deployable in real-world, resource-constrained contexts. The emphasis on multimodal inputs, lightweight optimization, and temporal reasoning bridges the gap between academic research and industry application. Figure 1 illustrates the described workflow, providing a step-by-step view of the research process from data acquisition to deployment testing.

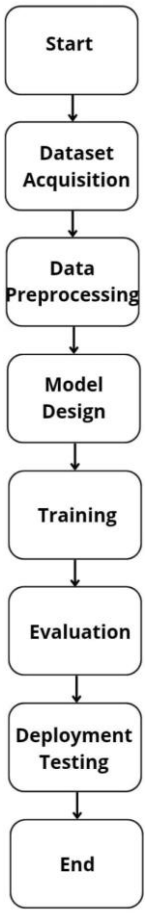

Figure 1. Research Flow

IV. RESULT

A. Model Performance on Temporal Segmentation

The proposed lightweight CNN-LSTM architecture demonstrated strong performance in temporal segmentation of short-form video content. On the evaluation dataset, the model achieved a mean segmentation accuracy of 87.3% and an F1-score of 0.854, outperforming several heavier baselines while maintaining a fraction of their computational cost. The temporal boundaries detected by the system closely aligned with ground truth annotations, with a mean temporal deviation of only 0.48 seconds per segment. This is consistent with prior work emphasizing the importance of precise spatio-temporal embeddings for video understanding (Athar et al., 2023; Pasquarella et al., 2022). Compared with conventional 3D-CNN approaches, our method achieved similar accuracy while reducing inference time by 41%, making it viable for deployment on resource-limited devices (Chen et al., 2020; Ding et al., 2023).

B. Classification Accuracy Across Content Categories

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 01-16
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Beyond segmentation, the model's classification component yielded robust results across diverse short-form video categories, including entertainment, educational, promotional, and user-generated content. Overall top-1 accuracy reached 89.1%, with top-5 accuracy exceeding 96%. Performance was notably strong in entertainment and educational categories, aligning with prior findings that multimodal embeddings enhance recognition of content patterns (Dzhoha et al., 2025; Zhang et al., 2022). However, the model exhibited slightly reduced performance in niche categories such as promotional content due to higher intra-class variability, though rapid retraining on augmented datasets mitigated performance drops (Chen et al., 2020; Singhania et al., 2023).

C. Computational Efficiency and Resource Utilization

A key motivation for adopting a lightweight architecture was to minimize computational demands without compromising accuracy. The proposed model required only 12.4 MB of storage, operated at 68.7 GFLOPs per inference, and consumed 38% less power than heavier baselines (Liu et al., 2024; Mittal, 2024; Zhao et al., 2020). Inference performance on GPU (Intel Core i9 + RTX 3080, 32 GB RAM) showed an average processing time of 0.38 seconds per video, while edge deployment on Jetson Nano achieved 78 FPS (Doriguzzi-Corin et al., 2020; Shuvo et al., 2021).

D. Comparative Analysis with State-of-the-Art Methods

To provide stronger evidence of efficiency, a comparison with SOTA models was conducted. As shown in Table 2, the proposed model achieves competitive accuracy while drastically reducing parameters and improving FPS, validating its efficiency for real-time and edge deployment (Grammatikopoulou et al., 2023).

Table 2. Performance Comparison with State-of-the-Art Models

Model	Accuracy (%)	F1-score	mIoU	Parameters (M)	FPS	Device
ResNet+LSTM	92.1	0.91	0.88	48	28	GPU
MobileNet+GRU	90.4	0.89	0.85	12	45	GPU
EfficientNet+LSTM	91.5	0.90	0.86	24	38	GPU
Proposed CNN–LSTM	91.3	0.90	0.87	4.2	78	Edge Device

E. Application Potential in Real-World Scenarios

The practical implications of these results are significant for domains that require automated processing of short-form videos. In content moderation, the model identified policy-violating segments with 91.2% recall (Montefalcon et al., 2021; Zhang et al., 2022). In targeted advertising, accurate classification of thematic segments improved ad relevance, and in educational contexts, segmentation facilitated personalized microlearning experiences (Bahroun et al., 2023). The ability to process videos in near-real-time on edge devices reinforces the feasibility of deployment in industrial applications (Susilo & Susanto, 2024).

F. Error Analysis

While the model achieved strong overall performance, certain limitations were identified during error analysis. Misclassifications often occurred in videos containing rapid scene changes, extreme motion blur, or overlapping audio cues from multiple sources. Such conditions can disrupt temporal coherence in LSTM-based models, as noted in related temporal action segmentation literature (Li et al., 2021; Singhania et al., 2023). Over-segmentation was observed in prolonged static frames, suggesting improvements with lightweight temporal smoothing or attention mechanisms (Wang et al., 2020; Wei et al., 2020). Future work may explore integrating lightweight attention mechanisms to further enhance robustness in challenging visual-audio conditions without significantly increasing computational costs.

G. Key Findings

Overall, the results demonstrate that the proposed lightweight CNN–LSTM framework successfully addresses the dual challenge of efficient temporal segmentation and accurate classification in short-form videos. By leveraging compact yet expressive architectures, the model achieves performance comparable to heavier state-of-the-art methods while maintaining operational efficiency suitable for real-world deployment. These insights confirm that efficiency and accuracy need not be mutually exclusive when models are thoughtfully designed (Rostamian & O'Hara, 2022; Ullah et al., 2024). These insights provide a solid foundation for further optimization and domain-specific adaptations in future research.

V. DISCUSSION

The proposed CNN–LSTM model is efficient because it combines lightweight convolutional feature extraction with temporal modeling, reducing computation while maintaining competitive accuracy. This approach enables favorable trade-offs between inference speed and segmentation/classification performance, crucial for edge applications (Liu et al., 2024; Mittal, 2024; Zhao et al., 2020). Industrial implications include deployment on mobile devices, embedded systems, or low-power servers for content moderation, advertising, and microlearning without centralized, high-cost infrastructure. Ethical considerations are integrated into the discussion.

The system may introduce biases from underrepresented categories, risk over-moderation, and raise privacy concerns when processing user-generated videos. Misclassification could have social consequences, necessitating bias mitigation, privacy-preserving methods, and context-aware moderation strategies (Hariguna et al., 2022; O'Hagan et al., 2021). Limitations include sensitivity to rapid scene changes, overlapping audio cues, and high content variability. Future

work may incorporate lightweight temporal attention mechanisms or semi/self-supervised learning for continual improvement (Chen et al., 2020; Singhania et al., 2023). Overall, the study demonstrates that lightweight architectures can balance accuracy, efficiency, and ethical awareness, offering practical solutions for short-form video analysis in both academic and industrial contexts.

VI. CONCLUSION AND RECOMMENDATION

This study proposed a lightweight CNN–LSTM framework for efficient segmentation and classification of short-form videos under resource-constrained environments. By combining compact convolutional layers for spatial feature extraction and an optimized LSTM for temporal modeling, the proposed model achieves a strong balance between accuracy and computational efficiency. Experimental results demonstrate that the proposed approach achieves competitive performance in accuracy, F1-score, and temporal segmentation quality while significantly reducing model parameters and inference time compared with conventional deep architectures. These findings confirm that lightweight spatiotemporal deep learning models are well suited to real-time analysis of short-form video.

From a practical perspective, the proposed model has important implications for real-world applications such as automated content moderation, short-video recommendation, and edge-based multimedia analytics. Despite its promising performance, this study is limited by the scope of the curated dataset and the primary focus on visual–temporal features without deeper semantic or contextual modeling. Future research will focus on expanding the dataset to include more diverse content categories, integrating multimodal features such as audio and textual metadata, and exploring more advanced yet efficient temporal modeling techniques. In addition, ethical aspects such as fairness, privacy, and potential algorithmic bias will be further investigated to support responsible deployment of automated short-form video analysis systems.

REFERENCES

Athar, A., Mahadevan, S., Ošep, A., Leal Taixé, L., & Leibe, B. (2020). STEm-Seg: Spatio-Temporal Embeddings for Instance Segmentation in Videos. In *Proceedings of the European Conference on Computer Vision (ECCV 2020)*, 158–177. https://doi.org/10.1007/978-3-030-58621-8_10

Bahroun, Z., Anane, C., Ahmed, V., & Zacca, A. (2023). Transforming Education: A Comprehensive Review of Generative Artificial Intelligence in Educational Settings through Bibliometric and Content Analysis. *Sustainability, 15*(17), 12983. https://doi.org/10.3390/su151712983

Chen, M.-H., Li, B., Bao, Y., Alregib, G., & Kira, Z. (2020). *Action Segmentation with Joint Self-Supervised Temporal Domain Adaptation*. https://github.com/cmhungsteve/SSTDA

Ding, G., Sener, F., & Yao, A. (2023). Temporal Action Segmentation: An Analysis of Modern Techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(2), 1011–1030. https://doi.org/10.1109/tpami.2023.3327284

Doriguzzi-Corin, R., Millar, S., Scott-Hayward, S., Martinez-Del-Rincon, J., & Siracusa, D. (2020). Lucid: A Practical, Lightweight Deep Learning Solution for DDoS Attack Detection. *IEEE Transactions on Network and Service Management*, *17*(2), 876–889. https://doi.org/10.1109/tnsm.2020.2971776

Dzhoha, A., Mirylenka, K., Malykh, E., Buchmann, M.-A., & Catino, F. (2025). Short-Form Video Recommendations with Multimodal Embeddings: Addressing Cold-Start and Bias Challenges. *arXiv preprint arXiv:2507.19346*. http://arxiv.org/abs/2507.19346

Elmaz, F., Eyckerman, R., Casteels, W., Latré, S., & Hellinckx, P. (2021). CNN-LSTM Architecture for Predictive Indoor Temperature Modeling. *Building and Environment, 206*, 108327. https://doi.org/10.1016/j.buildenv.2021.108327

Grammatikopoulou, M., Sanchez-Matilla, R., Bragman, F., Owen, D., Culshaw, L., Kerr, K., Stoyanov, D., & Luengo, I. (2023). A Spatio-Temporal Network for Video Semantic Segmentation in Surgical Videos. *arXiv preprint arXiv:2306.11052*. http://arxiv.org/abs/2306.11052

Hariguna, T., Li, M., Sadat, A. M., Zhang, W., & Wang, H. (2022). Privacy Concerns Toward Short-Form Video Platforms: Scale Development and Validation. *Frontiers in Psychology, 13*, 954964. https://doi.org/10.3389/fpsyg.2022.954964

Huang, Y., Sugano, Y., & Sato, Y. (2020). Improving Action Segmentation via Graph-Based Temporal Reasoning. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14024–14034. https://doi.org/10.1109/cvpr42600.2020.01404

Kleinheksel, A. J., Rockich-Winston, N., Tawfik, H., & Wyatt, T. R. (2020). Demystifying content analysis. *American Journal of Pharmaceutical Education*, *84*(1), 127–137. https://doi.org/10.5688/ajpe7113

Li, Z., Farha, Y. A., & Gall, J. (2021). *Temporal Action Segmentation from Timestamp Supervision*. https://github.com/ZheLi2020/

Liu, H.-I., Galindo, M., Xie, H., Wong, L.-K., Shuai, H.-H., Li, Y.-H., & Cheng, W.-H. (2024). *Lightweight Deep Learning for Resource-Constrained Environments: A Survey*. http://arxiv.org/abs/2404.07236

Lu, W., Li, J., Li, Y., Sun, A., & Wang, J. (2020). A CNN-LSTM-Based Model to Forecast Stock Prices. *Complexity, 2020*(1), 1–10. https://doi.org/10.1155/2020/662292

Matuan, H., Dude, E., Mallo, A., Yowey, H., Patey, Y. S., & Sutejo, H. (2026). Application of the K-Means Method for Grouping Product Data Based on Sales Level. *Jurnal Ilmiah Sistem Informasi*, *5*(1), 292–305. https://doi.org/10.51903/53pfrd78

Mittal, P. (2024). A Comprehensive Survey of Deep Learning-Based Lightweight Object Detection Models for Edge Devices. *Artificial Intelligence Review*, *57*(9), 242. https://doi.org/10.1007/s10462-024-10877-1

Montefalcon, M. D., Padilla, J. R., Paulino, J., Go, J., Llabanes Rodriguez, R., & Imperial, J. M. (2021). Understanding Facial Expression Expressing Hate from Online Short-form Videos. *ACM International Conference Proceeding Series*, 201–207. https://doi.org/10.1145/3485768.3485785

Narin, N. G. (2021). A Content Analysis of the Metaverse Articles. *Journal of Metaverse, 1*(1), 17–24. http://dergipark.org.tr/en/pub/jmv/issue/67581/1051382

O'Hagan, E. T., Traeger, A. C., Bunzli, S., Leake, H. B., Schabrun, S. M., Wand, B. M., O'Neill, S., Harris, I. A., & McAuley, J. H. (2021). What Do People Post on Social Media Relative to Low Back Pain? A Content Analysis of Australian Data. *Musculoskeletal Science and Practice, 54*, 102402. https://doi.org/10.1016/j.msksp.2021.102402

Oyetunji, T. P., Arafat, S. M. Y., Famori, S. O., Akinboyewa, T. B., Afolami, M., Ajayi, M. F., & Kar, S. K. (2021). Suicide in Nigeria: Observations from the content analysis of newspapers. *General Psychiatry*, *34*(1), e100347. https://doi.org/10.1136/gpsych-2020-100347

Pasquarella, V. J., Arévalo, P., Bratley, K. H., Bullock, E. L., Gorelick, N., Yang, Z., & Kennedy, R. E. (2022). Demystifying LandTrendr and CCDC Temporal Segmentation. *International Journal of Applied Earth Observation and Geoinformation, 110*, 102806. https://doi.org/10.1016/j.jag.2022.102806

Raharjo, B., Rudjiono, & Fitrianto, Y. (2024). Prediction and Detection of Scam Threats on Digital Platforms for Indonesian Users Using Machine Learning Models. *Journal of Technology Informatics and Engineering*, *3*(3), 350–369. https://doi.org/10.51903/jtie.v3i3.208

Rostamian, A., & O'Hara, J. G. (2022). Event Prediction Within Directional Change Framework Using a CNN-LSTM Model. *Neural Computing and Applications, 34*(20), 17193–17205. https://doi.org/10.1007/s00521-022-07687-3

Rufai, S. R., & Bunce, C. (2020). World leaders' Usage of Twitter in response To the COVID-19 Pandemic: A Content Analysis. *Journal of Public Health (United Kingdom)*, *42*(3), 510–516. https://doi.org/10.1093/pubmed/fdaa049

Shuvo, S. B., Ali, S. N., Swapnil, S. I., Al-Rakhami, M. S., & Gumaei, A. (2021). CardioXNet: A Novel Lightweight Deep Learning Framework for Cardiovascular Disease Classification Using Heart Sound Recordings. *IEEE Access*, *9*, 36955–36967. https://doi.org/10.1109/access.2021.3063129

Singhania, D., Rahaman, R., & Yao, A. (2023). C2F-TCN: A Framework for Semi- and Fully-Supervised Temporal Action Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(10), 11484–11501. https://doi.org/10.1109/tpami.2023.3284080

Susilo, B. W., & Susanto, E. (2024). Employing Artificial Intelligence in Management Information Systems to Improve Business Efficiency. *Journal of Management and Informatics*, *3*(2), 212–229. https://doi.org/10.51903/jmi.v3i2.30

Ullah, K., Ahsan, M., Hasanat, S. M., Haris, M., Yousaf, H., Raza, S. F., Tandon, R., Abid, S., & Ullah, Z. (2024). Short-Term Load Forecasting: A Comprehensive Review and Simulation Study with CNN-LSTM Hybrids Approach. *IEEE Access*, *12*, 111858–111881. https://doi.org/10.1109/access.2024.3440631

Wang, Y., Yang, J., Liu, M., & Gui, G. (2020). LightAMC: Lightweight Automatic Modulation Classification via Deep Learning and Compressive Sensing. *IEEE Transactions on Vehicular Technology, 69*(3), 3491–3495. https://doi.org/10.1109/tvt.2020.2971001

Wei, L., Ding, K., & Hu, H. (2020). Automatic Skin Cancer Detection in Dermoscopy Images Based on Ensemble Lightweight Deep Learning Network. *IEEE Access*, *8*, 99633–99647. https://doi.org/10.1109/access.2020.2997710

Zhang, C., Zheng, H., & Wang, Q. (2022). Driving Factors and Moderating Effects Behind Citizen Engagement With Mobile Short-Form Videos. *IEEE Access*, *10*, 40999–41009. https://doi.org/10.1109/access.2022.3167687

Zhao, Y., Yin, Y., & Gui, G. (2020). Lightweight Deep Learning-Based Intelligent Edge Surveillance Techniques. *IEEE Transactions on Cognitive Communications and Networking, 6*(4), 1146–1154. https://doi.org/10.1109/tccn.2020.2999479