IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 February 2024, accepted 22 March 2024, date of publication 29 March 2024, date of current version 4 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3383060

RESEARCH ARTICLE

The Short Video Popularity Prediction Using Internet of Things and Deep Learning

ZICHEN HE[1] AND DANIAN LI[2]
[1]School of Journalism and Communication, Chongqing Normal University, Chongqing 401331, China
[2]School of Electronics and Internet of Things, Chongqing College of Electronic Engineering, Chongqing 401331, China
Corresponding author: Zichen He (20131554@cqnu.edu.cn)

ABSTRACT In order to furnish valuable insights and solutions applicable to content creators, social media platforms, academic research, and general users, this investigation integrates the Internet of Things (IoT) with deep learning regression models to examine methodologies for predicting the popularity of short videos. Within the context of cross-cultural communication, a proposed Content Popularity Rank Prediction based on the Convolutional Neural Network (CPRP-CNN) model relies exclusively on the personal attributes of the publisher and the textual characteristics of short videos to anticipate the viewership levels of short videos promptly following their release. Through simulated experiments, the model's performance is assessed, revealing that the utilization of the Rectified Linear Unit (Relu) activation function in the CPRP-CNN model enhances accuracy by 42.2% when contrasted with the use of the sigmoid function. This enhancement is coupled with a 37.8% reduction in cross-entropy loss. Furthermore, the proposed CPRP-CNN model attains a cross-entropy of 0.692 and an accuracy of 74.7%, exhibiting superior Mean Squared Error (MSE) and Mean Absolute Error (MAE) values of 2.728 and 1.751, respectively, when compared to alternative prediction models. These outcomes signify that the amalgamation of deep learning models with fused features within the IoT context significantly ameliorates the predictive efficacy of short video popularity. The research findings contribute to the enhancement of personalized and precise short video content recommendations.

INDEX TERMS Cross-cultural communication, Internet of Things, deep learning regression model, short video, popularity prediction.

I. INTRODUCTION

A. RESEARCH BACKGROUND AND MOTIVATIONS

In recent years, there has been a global surge in the popularity of short videos, propelled by applications such as TikTok, Instagram, and Snapchat. These platforms have emerged as primary social media channels, particularly for the younger generation, notably Generation Z, amassing billions of users [1], [2], [3]. These applications for short videos not only furnish a straightforward and engaging avenue for sharing snippets of life and generating content but also cultivate a novel social culture, facilitating user communication and interaction on a global scale. However, the popularity and communication dynamics of short videos are influenced by various factors, including user interests, content quality, social network effects, geographical location, and others. The intricate interplay of these factors poses a formidable challenge in accurately predicting the popularity of short videos [4], [5], [6]. Conventional approaches often hinge on uncomplicated metrics like user feedback, viewing frequencies, or likes; nevertheless, these metrics fall short of fully capturing the intricacies inherent in short video communication.

In this context, research on a regression model that integrates the Internet of Things (IoT) and deep learning becomes crucial [7], [8]. IoT technology enables the collection of comprehensive and detailed data, encompassing environmental information during users' short video consumption—such as location, device type, and network connection quality. This additional information enhances the understanding of the user context, leading to more precise predictions of short video popularity [9], [10], [11]. Concurrently, deep

The associate editor coordinating the review of this manuscript and approving it for publication was Vlad Diaconita.

learning regression models exhibit robust data modeling capabilities, handling large-scale, high-dimensional data and discerning intricate patterns and relationships [12], [13], [14]. Combining IoT data with deep learning regression models facilitates the establishment of a more accurate prediction model, shedding light on short video propagation mechanisms and influencing factors [15], [16], [17]. Nevertheless, managing the vast, complex, and diverse data generated by the IoT poses significant challenges. Effectively processing, analyzing, and extracting valuable information from such data requires advanced techniques. Additionally, uncertainties and noise in data may adversely impact prediction model performance. The IoT's diverse device landscape raises compatibility and standardization issues, necessitating seamless connection and data-sharing solutions—a pivotal challenge in short video popularity prediction within the IoT.

This study's motivation stems from the considerable commercial value associated with accurately predicting short video popularity for content creators and social media platforms. Anticipating which videos are likely to gain traction allows for strategic content planning, attracting more users and advertisers, and ultimately augmenting revenue. Furthermore, understanding user behavior variations across regions and cultures facilitates personalized content recommendations, enhancing user satisfaction. The research in this domain contributes to a profound comprehension of information dissemination and social interaction similarities and differences across diverse cultural backgrounds, uncovering the potential applications of IoT technology and deep learning in cross-cultural communication research.

B. RESEARCH OBJECTIVES

The primary objective of this investigation is to delineate a method for predicting the popularity of short videos within the context of cross-cultural communication. This study seeks to integrate the Internet of Things (IoT) with a deep learning regression model with the objective of providing valuable insights and solutions for content creators, social media platforms, academic research, and the wider user community.

In Section I, the study expounds on the background, objectives, and imperative nature of the research. Section II provides a comprehensive review of the existing research landscape encompassing deep learning regression models, IoT-related technologies, and methodologies pertinent to predicting short video popularity. Section III, guided by the framework of cross-cultural communication and leveraging the IoT backdrop, introduces a short video popularity prediction model rooted in the deep learning regression paradigm. This model incorporates a multi-modal feature fusion supervision approach. Section IV is dedicated to scrutinizing and affirming the effectiveness of the proposed prediction model. Finally, Section V encapsulates the study's contributions, acknowledges its limitations, and delineates potential avenues for future research.

II. LITERATURE REVIEW

The profound success of deep learning regression models in diverse domains, spanning image processing, natural language processing, and recommendation systems, is noteworthy. In the realm of short video popularity prediction, numerous studies have leveraged deep learning models to enhance predictive accuracy. Din et al. [18] introduced a method for short video popularity prediction based on the Convolutional Neural Network (CNN). Employing an extensive dataset of short videos, they utilized the CNN model to extract visual features, subsequently merging these features with social network data to forecast popularity. Outcomes underscored that the deep learning model adeptly captures intricate relationships between visual and social information, thereby elevating prediction accuracy. Similarly, Waqas et al. [19] proposed a short video popularity prediction approach founded on Recurrent Neural Network (RNN). Focusing on user behavior sequences during short video consumption, they employed the RNN model to encode these sequences, revealing superior predictive performance when accommodating dynamic shifts in user behavior.

Simultaneously, advancements in IoT technologies have revolutionized data collection and transmission capabilities. These innovations yield copious amounts of environmental data, offering pivotal insights for short video popularity prediction. Abidi et al. [20] explored the integration of IoT technology for social media data collection. They devised a data acquisition system predicated on sensors and smart devices, facilitating the collection of environmental information during short video consumption, encompassing factors such as temperature, humidity, and illumination. This dataset proved instrumental in analyzing user viewing experiences and short video dissemination patterns. Additionally, Liu et al. [21] delved into the application of IoT technology for user behavior data collection. Leveraging IoT sensors to track user movements and gestures during short video consumption, they examined user emotions and engagement levels. These insights proved invaluable in comprehending user reactions to short videos and forecasting their popularity.

The domain of short video popularity prediction constitutes a multifaceted research area encompassing diverse methodologies and technologies. Alternative approaches have been explored in this realm beyond the utilization of deep learning regression models and IoT technology. Abbas et al. [22] introduced a method for short video popularity prediction grounded in social network analysis. Through an examination of social network relationships among users, they constructed a user influence propagation model, facilitating the prediction of propagation trajectories and the ensuing popularity of short videos. Furthermore, Brito and Adeodato [23] investigated a short video popularity prediction method centered on text analysis. Leveraging natural language processing technology to scrutinize the titles, descriptions, and user comments associated with short videos, they extracted text features to prognosticate short video popularity. Table 1 summarizes current research status.

TABLE 1. Summary of current research status.

Researcher	Method	Technical Details	Dataset	Results
Din et al. (2019)	Deep learning model based on CNN	Extracting video visual features using CNN, combined with social network data	Large-scale short video dataset	Improved prediction accuracy
Waqas et al. (2022)	Deep learning model based on RNN	Modeling user behavior sequences using RNN	Short video dataset	Considering dynamic changes in user behavior, enhanced prediction accuracy
Abidi et al. (2020)	IoT technology data collection	Collecting environmental data using sensors and smart devices	User watching dataset	Analyzing user viewing experience and short video propagation
Liu et al. (2019)	IoT technology data collection	Tracking user movements and gestures using IoT sensors	User watching dataset	Studying user emotion and engagement, predicting popularity
Abbas et al. (2023)	Method based on social network analysis	Analyzing social network relationships among users, constructing user influence propagation model	Social media dataset	Predicting short video propagation paths and popularity
Brito et al. (2023)	Method based on text analysis	Analyzing text features using natural language processing technology	Short video titles, descriptions, and comments dataset	Predicting short video popularity

In summary, the ongoing evolution and integration of established research methodologies contribute to an enhanced and more precise prediction of short video popularity, thereby advancing the progress of digital media and social interaction. This study introduces a short video popularity prediction model system based on deep learning regression models, considering multi-modal feature fusion supervision modeling and thoroughly accounting for the interactions between features. By leveraging IoT data, this research provides additional dimensions of information for short video prediction.

III. RESEARCH MODEL

A. THE POPULARITY OF SHORT VIDEOS FROM THE PERSPECTIVE OF CROSS-CULTURAL COMMUNICATION

Cultural disparity stands as a significant determinant influencing the popularity of short videos. Divergent cultural backgrounds engender distinct perceptions and values regarding humor, emotional expression, and moral concepts, thereby shaping the audience's comprehension and acceptance of short video content. The appropriateness of self-mockery and satire, for instance, may vary across cultures, being considered impolite or offensive in some and perceived as humorous expressions in others. Consequently, creators of short videos must contemplate the acceptability among diverse cultural audiences to ensure broad dissemination during cross-cultural communication.

Moreover, the content attributes of short videos play a pivotal role in their popularity within cross-cultural communication. Varied thematic focuses, such as emotional expression or humor and entertainment, cater to diverse viewer preferences. Producers of short videos are thus tasked with selecting content characteristics aligned with the cultural background and preferences of the target audience. Additionally, the incorporation of multilingual subtitles and localized content serves to enhance the appeal of short videos in cross-cultural contexts. Social media platforms and video-sharing applications have emerged as principal channels for disseminating short videos. Nevertheless, the usage patterns and regulations governing social media differ across cultural domains. Consequently, short video producers must comprehend the distinctive features of various social media platforms, enabling them to formulate tailored communication strategies and enhance the visibility and communicative impact of their short videos.

B. SHORT VIDEO CACHE STRATEGY BASED ON IoT TECHNOLOGY

Presently, the proliferation of micro-base stations contributes to an augmented network density, progressively diminishing the jurisdiction of individual micro-base stations [24], [25], [26]. Although this expansion curtails coverage areas, allowing networks to encompass more users, it concurrently leads to the shortened proximity between micro-base stations. Consequently, mobile users undergo frequent handovers as they traverse distinct coverage regions of micro-base stations.

Illustrated in Figure 1 is the influence of user transitions on caching services. Consider a user's movement as an instance, where the user is presently linked to base station 1 and will later traverse the path {1,2,3}. During this trajectory, the user initiates a file service request with base station 1. However, an intervening base station switch occurs during the file transmission process, rendering base station 1 incapable of fulfilling the user's request. This is attributed to the absence of the required files in the caches of base stations 2 and 3, compelling users to invest additional time in content retrieval from the core network. Such switching operations amplify the temporal costs incurred by users.

Alternatively, when transitioning between base stations, such as from base station 1 to base station 2, immediate access to the requisite files becomes feasible, ensuring the seamless culmination of the entire service process without additional waiting periods. Addressing this challenge necessitates the

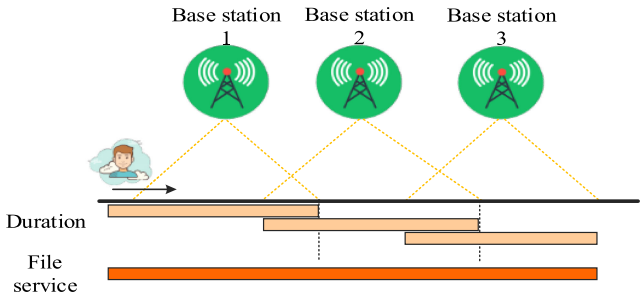

FIGURE 1. Impact of user switching on caching service.

formulation of pertinent user mobility handling protocols, mitigating the supplementary waiting time costs for users and enhancing their overall experience.

Short videos exhibit a characteristic wherein users' business requests undergo a rapid refresh, predominantly attributed to the widespread prevalence of diverse short videos. Notably, the highly popular short videos constitute nearly 80% of the total, while the majority of other short videos demonstrate comparatively lower popularity [27], [28]. Moreover, the request frequency for most short video services tends to be relatively modest, with numerous requests concentrated on specific highly popular files. Noteworthy is the pattern where many of these requests transpire singularly, occurring only once, after which the same files are not subject to subsequent requests.

Language models can be employed to process textual information within videos, such as titles, descriptions, and user comments [29]. Through the utilization of natural language processing techniques, language models can extract textual features from videos and comprehend the semantic information embedded within them. This contributes to a better understanding of the video's theme, emotions, and content quality. Visual Language Models (VLM) are designed to handle the visual information within videos, encompassing aspects like color, shape, motion, and scenes. Leveraging computer vision technology and deep learning algorithms, VLM can extract crucial features from videos, identifying key elements and scenes. This facilitates the comprehension of the visual effects and attractiveness of the video.

C. MULTI-MODAL DATA FUSION MODE

Multi-modal data fusion involves the integration of information derived from various perceptual modes, such as text, image, audio, etc., into a comprehensive model or system aimed at enhancing the capacity for understanding and analyzing data [30], [31], [32]. The primary objective of multi-modal fusion is to enhance the complementarity among modes, allowing each mode to furnish additional information to augment the overall analytical performance.

Hybrid fusion encompasses both late fusion and early fusion, amalgamating post and telecommunications. By mapping the original input data into their high-level

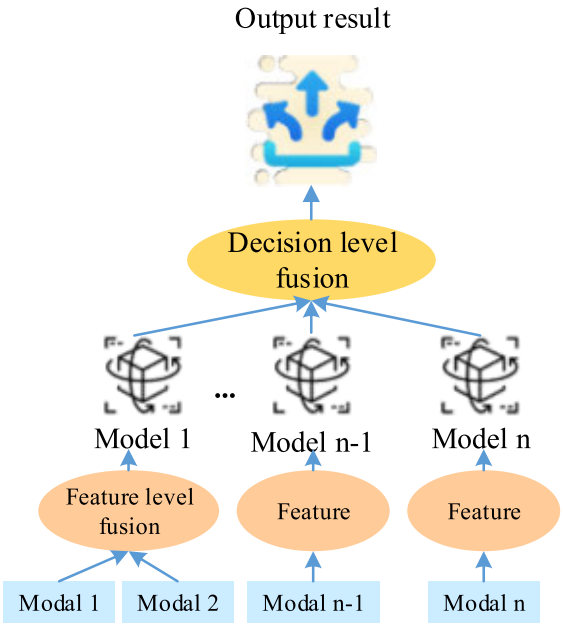

FIGURE 2. Schematic diagram of hybrid fusion method.

representations, the fusion of diverse modal information and the acquisition of cross-modal representations become more achievable [33], [34], [35]. This approach facilitates modal fusion through both early and late fusion at distinct fusion stages. As deep learning models demonstrate flexibility and diversity advantages, the field of deep multi-modal fusion is also evolving. In terms of implementation methods, deep multi-modal fusion typically adheres to the principles of hybrid fusion [36], [37], [38]. This is attributed to the hybrid fusion method exhibiting minimal information loss and reduced fusion complexity, enabling greater flexibility in customizing fusion methods during feature extraction and decision-making processes. Figure 2 illustrates a schematic diagram of the hybrid fusion method.

D. CONTENT POPULARITY PREDICTION BASED ON MULTI-MODAL FUSION REGRESSION MODEL

This study employs a diverse set of factors to forecast the forthcoming popularity of short videos, encompassing early viewing times, textual attributes, publisher characteristics, and user interaction information. Initially, the popularity of short video content is delineated, typically quantified by the number of views, and denoted as $x_t = \{x_0, x_1, \ldots, x_t\}$. Through the collection of data within the initial n time intervals, the prediction of content popularity in the subsequent n time intervals becomes feasible. For any video posted on a short video website, the start time of observation is defined as $T_0 = t - n\Delta t$ to predict the future content popularity, which is defined as shown in Eq, (1):

$$y_{(t+n\Delta t)} = \begin{cases} f\left(U_t, tx\right), & t < \Delta t \& n = 1 \\ g\left(x_t, S_t, U_t, tx\right), & otherwise \end{cases} \quad (1)$$

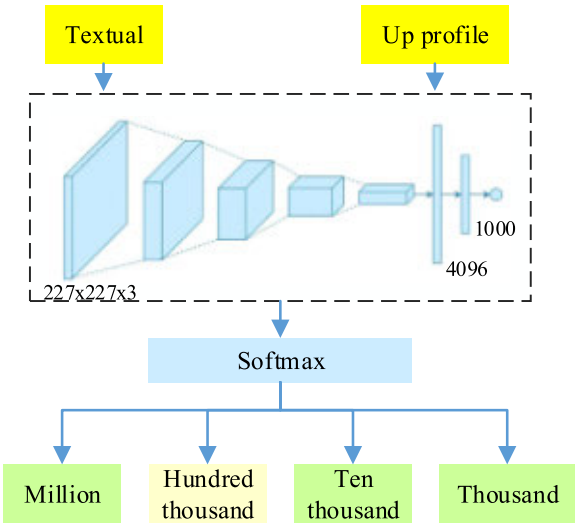

FIGURE 3. Structure of content popularity classification prediction model based on CNN.

In Eq. (1), $t < \Delta t$ indicates that the content has just been released, there is not enough historical data, including broadcast volume and social characteristics, to accurately predict the popularity of future content. Therefore, in this case, the problem of content popularity prediction can be transformed into a classification problem, and the short video publisher's characteristics and text characteristics collected at time t can be used to predict which level of the video's broadcast volume at time $t + \Delta t$ belongs to.

The dissemination of short videos occurs in a stochastic manner, garnering a substantial user base upon release [39], [40], [41], [42]. Enhancing network performance and user experience hinges on accurately predicting the viewership volume during the initial release phase of short videos. Typically, when a content creator releases a short video, access to historical viewership data and social attributes of the video is unattainable. However, textual characteristics and personal attributes of the content creator are readily available [43], [44], [45]. In order to address this challenge, this study introduces a Content Popularity Rank Prediction based on Convolutional Neural Network (CPRP-CNN). This model relies solely on the personal attributes of the content creator and the textual features of short videos to predict the viewership volume level at the point of release. The architecture of the content popularity classification prediction model based on CNN is depicted in Figure 3.

The calculation of the CPRPC model is shown in Eq. (2):

$$y_{(t+n\Delta t)} = f\left(U_t, tx\right) = \frac{e^{o_{c,i}}}{\sum_{i=1}^{d} e^{o_{c,i}}} \tag{2}$$

In Eq. (2), $o_{c,i}$ represents the output of CNN and d is the number of neurons in the Softmax layer.

In the output of the convolution layer, the k-th convolution kernel W traverses the whole input $O(0)$ and generates a feature vector $O(1,k)$, which is expressed as shown in Eq. (3):

$$O_{i,j}^{(1,k)} = R\left(\sum_{t=0}^{r_k-1} W_{s,t}^{(1,k)} \cdot O_{i+s,j+t}^{(0)} + b^{(1,k)}\right) \tag{3}$$

In Eq. (3), s represents the stride, r_k is the k convolution kernel, and $b^{(1,k)}$ is the offset.

The maximum pooling calculation can be expressed as shown in Eq. (4):

$$O_{i,j}^{(2,k)} = Maxpolling\left(\sum_{t=0}^{r_k} O_{i+s,j+t}^{(1)}\right) \tag{4}$$

The step of the CPRP-CNN algorithm is shown in Table 2.

TABLE 2. Steps of CPRP-NN algorithm.

Input: Short video data set D_{train}
Output: Predict the short video tag of the data set
1. Process multi-modal data
2. Import the CPRP-CNN model and set the Batch size
3. Load model pre-training weights
4. For i=1:Max_Epoch:
5. Read a small batch of data $\{(O^1, y^1), (O^2, y^2), ...\}$
6. Define multi-modal features ReSize
7. Input the multi-modal features of each video in the dataset into the CPRP-CNN model
8. Take out the prediction tag corresponding to each short video and save it in batches
9. End

In the CPRP-CNN model, convolutional layers are employed to extract features from the textual content of short videos. By configuring various convolutional kernel sizes and strides, the model can extract different phrases and semantic information from the text. These features are then fed into pooling layers for dimensionality reduction, aiming to enhance the model's generalization ability by reducing feature dimensions. In addition to textual features, personal attributes of the content creator, such as historical viewership and follower count, are taken into account. These features are fused with textual features through a fully connected layer, providing a comprehensive consideration of both the content quality of short videos and the impact of the content creator's influence on their popularity.

Within the CPRP-CNN model, the Rectified Linear Unit (ReLU) activation function is employed. The ReLU activation function effectively addresses the vanishing gradient problem, enhancing the model's training speed and stability. Compared to the sigmoid activation function, ReLU demonstrates better performance when training deep neural networks. Mean Squared Error (MSE) is utilized as the loss

function, and model parameters are optimized through the Adam optimizer. The Adam optimizer is an adaptive learning rate optimization algorithm that dynamically adjusts the learning rate based on the model's training progress, resulting in improved convergence effects.

Prior to initiating model training, a suitable weight initialization method is applied, and bias values are initialized for each neuron and layer.

Throughout the training process:

1. Forward Propagation: Input data is conveyed through the model, and the predicted value is computed.

2. Loss Calculation: The loss is computed based on the predicted value and the actual value.

3. Back Propagation: Gradients are computed in relation to the loss, and the weights and biases are updated utilizing the gradient descent algorithm.

4. Learning Rate Scheduling: The learning rate is adjusted in accordance with the progression of training to mitigate fluctuations in the optimization process.

5. Model Saving/Loading: During training, the model's weights and parameters can be periodically saved for subsequent loading and utilization as needed.

In the verification/testing process:

1. Verification Dataset Evaluation: The verification dataset is utilized to assess the model's performance, facilitating parameter adjustments or model selection.

2. Test Dataset Evaluation: The test dataset is employed to evaluate the ultimate performance of the model, ensuring that overfitting has not occurred.

3. Model Evaluation Indicators: Appropriate evaluation metrics are chosen based on the task type, such as accuracy, recall, F1 score, etc.

4. Model Optimization: The model is further optimized or adjusted based on the outcomes of the verification/test results.

When evaluating the most crucial harmonics in videos, an assessment method based on the signature transform benchmark can be introduced [46]. The signature transform is a signal processing technique that extracts specific features from a signal, generating a distinctive ''signature.'' In video analysis, the signature transform can be employed to identify and extract key harmonics in the video, thereby assessing the quality and significance of the video.

IV. EXPERIMENTAL DESIGN AND PERFORMANCE EVALUATION

A. DATASETS COLLECTION

The dataset utilized for short video popularity prediction in this investigation was curated by the Media Laboratory of the National University of Singapore (http://acmmm2016.wixsite.com/micro-videos). Comprising 303,242 user-generated short videos, the dataset was sourced from the Vine platform, an online short video-sharing website, and contributed by 98,166 distinct users. The data for this study was obtained through legal means, and explicit authorization and consent were obtained from users. During the collection of user behavioral data, the data underwent anonymization to safeguard the personal privacy of users.

Given the intrinsic connection between popularity and online social interaction, the computation of the ultimate popularity score for short videos necessitates carefully considering four key metrics—namely, the count of comments, reposts, likes, and view cycles. These metrics are averaged and normalized to ensure the resultant score falls within the range of 0 to 1.

For experimentation purposes, a dataset comprising 8810 entries is employed, and multiple rounds of randomized experiments are executed. Each experimental round utilizes 90% of the samples for training, with the remaining 10% reserved for testing. The final experimental outcome is derived by averaging the results from 10 test iterations.

B. EXPERIMENTAL ENVIRONMENT

Within the scope of this investigation, the training and testing procedures are executed utilizing TensorFlow GPU, with the graphics card specified as GeForce RTX 3090. The operational environment of the model framework adheres to the following specifications: python=3.8 and pytorch=1.7.1.

C. PARAMETERS SETTING

The specific hyperparameter settings of the model are shown in Table 3.

TABLE 3. Parameter setting of CPRP-CNN model training process.

Hyperparameter	description	Numerical value
Input size	The image contains channels, all of which are 35x36 matrices	35×36×2
CNN layers	Convolution layer, ReLU layer, pooling layer, and fully connected layer	6
Output size	The output unit includes four categories	4
Batch size	Number of data input to the model at a time	64
Learning rate	Confirmed the weight update speed	0.001
optimizer	Update parameters during model training	Adam
Epoch	-	200

D. PERFORMANCE EVALUATION

1) INFLUENCE OF DIFFERENT ACTIVATION FUNCTIONS ON THE PERFORMANCE OF CLASSIFICATION PREDICTION MODEL

During the training phase of the CPRPC model, the meticulous selection of internal parameters holds paramount significance, contributing to the effective mitigation of common issues such as model overfitting and gradient vanishing. The activation function assumes a pivotal role in neural networks, primarily facilitating the nonlinear transformation of data to

address the limitations of linear models in handling classification tasks. The discernible observation from Figure 4 indicates that the Sigmoid activation function encounters challenges related to gradient saturation within deep neural networks featuring a multi-layer structure. In such instances, during the backpropagation process, the derivative of the Sigmoid function tends to approach zero, resulting in sluggish weight update rates and ultimately impeding the attainment of a satisfactory level of model accuracy.

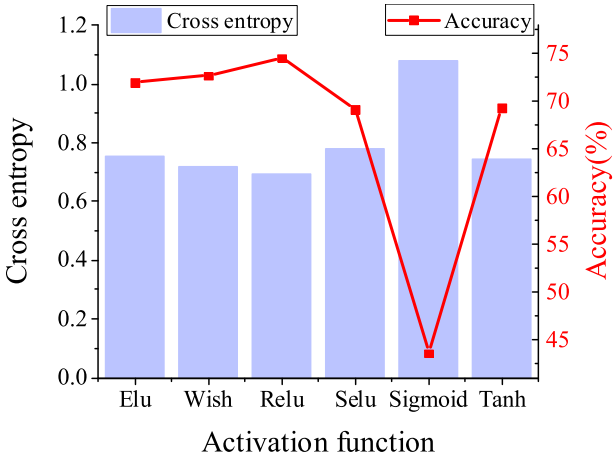

FIGURE 4. Influence of different activation functions on the performance of classification prediction model.

In contrast, the Relu activation function adeptly addresses this issue. Compared to the Sigmoid function, the Relu activation function within the CPRP-CNN model enhances model accuracy by 42.2% and reduces cross-entropy loss by 37.8%. This improvement is attributed to the issues of gradient vanishing or exploding that arise with the Sigmoid function when dealing with excessively large or small inputs, leading to challenges in model training and convergence. In contrast, the Relu activation function exhibits superior properties, effectively alleviating the problems of gradient vanishing and exploding while concurrently expediting the model's training process.

In order to assess the efficacy of the CPRP-CNN classification prediction model proposed in this study, a comparative analysis is conducted with Multi-Layer Perceptron (MLP), traditional CNN Lenet5, and AlexNet. The performance metrics considered encompass cross-entropy and accuracy. As depicted in Figure 5, both CNN and AlexNet outperform MLP in terms of accuracy. Notably, the proposed CPRP-CNN attains a cross-entropy score of 0.692 and achieves an accuracy level of 74.7%.

2) COMPARISON OF DIFFERENT POPULARITY PREDICTION METHODS

A comprehensive comparison of MSE and Mean Absolute Error (MAE) among various popularity prediction models is undertaken, as depicted in Figure 6. The popularity prediction approach proposed in this study, CPRP-CNN,

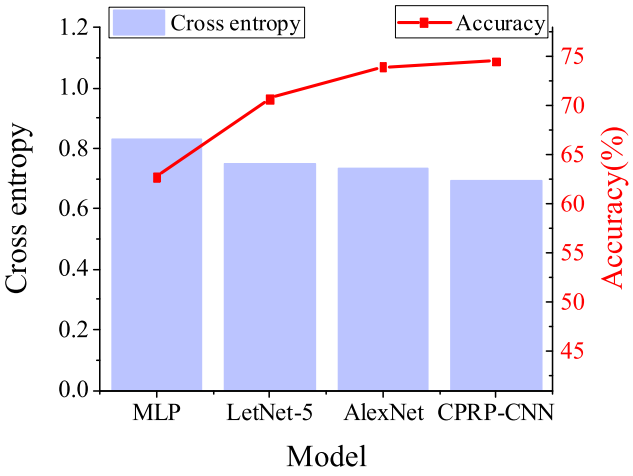

FIGURE 5. Performance comparison of different classification prediction models.

demonstrates superior predictive performance, yielding an MSE of 2.728 and an MAE of 1.751. Figure 7 exhibits the outcomes of the ablation study. In this investigation, multiple modalities are considered for the popularity prediction task, encompassing images, image attributes, text, and explicit attributes. These modalities play a pivotal role in the process of multi-modal feature fusion, signifying their paramount importance in predicting content popularity.

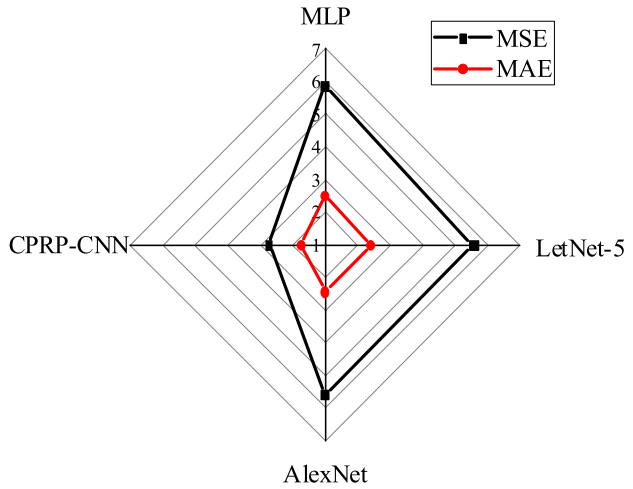

FIGURE 6. Performance comparison of different classification prediction models.

E. DISCUSSION

This study endeavors to comprehend the global popularity of short videos by scrutinizing user behavior and social interaction across diverse cultural backgrounds [47], [48]. The study uses a multi-modal feature fusion methodology encompassing images, texts, and user behavior data to capture users' preferences and interactions with short videos within distinct cultural environments. This approach proves highly successful in predicting popularity, surpassing the

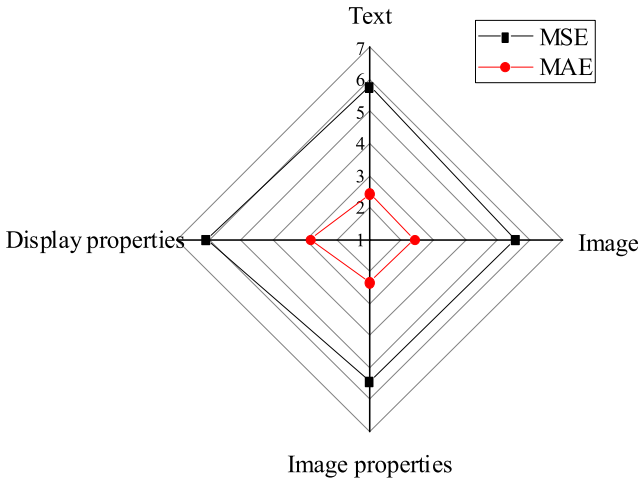

FIGURE 7. Experimental results of ablation research.

limitations of traditional methods like linear regression and logistic regression, which often struggle to grasp intricate user behaviors and social patterns. The multi-modal feature fusion method adopted here enhances prediction accuracy by comprehensively integrating images, texts, and user behavior data, outperforming single-model approaches.

Ji et al. [49] identified two categories of fusion methods: multi-factor fusion and multi-model fusion. While multi-factor fusion was straightforward and easily implemented, it might have overlooked correlations between diverse factors due to its reliance on simple concatenation operations. Conversely, multi-model fusion excelled at integrating various factors but may have neglected interactive information between them, presenting implementation challenges. Practical application necessitated designing the fusion mode based on specific requirements. As emphasized by Hartono [50], algorithms for predicting popularity that leverages multi-modal features comprehensively consider user and social characteristics. These algorithms employ machine learning models to forecast the popularity of content. Ongoing enhancements to these models elevated overall predictive performance, although they often demanded substantial data and computing resources. In contrast, the multi-modal feature fusion method exhibited commendable performance on smaller datasets and was more accessible for implementation.

In instances of limited data availability, intricate data patterns may pose challenges for the model to capture effectively. In order to address this predicament, certain scholarly inquiries have introduced approaches involving the utilization of low-complexity models or sparse representation. These methodologies aim to curtail the model's parameter count, thereby diminishing reliance on data and augmenting the interpretability of the model. In scenarios marked by data scarcity, the importance of feature selection becomes pronounced. Opting for features most pertinent to the target task serves to mitigate noise and redundancy in the data, consequently enhancing the model's performance. Furthermore, select research endeavors have introduced feature selection methods grounded in sparse representation, enabling the automatic identification of features most germane to the target task and thereby further refining the model's performance. Video quality assessment was a crucial aspect in evaluating the content quality of videos, and spatiotemporal modeling stood out as a key technology within this domain. Fang et al. [51] asserted that spatiotemporal modeling aimed to capture dynamic temporal and spatial features in videos, enabling a comprehensive evaluation of video quality. By establishing temporal and spatial models, it became possible to extract and compare features for each frame, thus providing a more accurate assessment of video quality. This study discussed the use of deep learning models and IoT technology to extract visual features and user behavior data from videos, both of which could be utilized in spatiotemporal modeling. Wang et al. [52] proposed that mobile 360-degree video streaming was a novel video format offering a panoramic perspective and immersive viewing experience. However, due to its unique shooting and playback methods, assessing the quality of mobile 360-degree video streams presented significant challenges. In order to enhance the evaluation of the quality of mobile 360-degree video streams, there is a need to explore significance-driven quality adaptation for this format.

In practical applications, the stages of data collection and preprocessing assume paramount significance. Ensuring the quality of collected data, including its accuracy and completeness, is imperative. Simultaneously, meticulous preprocessing of the data, encompassing cleaning and standardization, becomes crucial to mitigate the impact of noise and aberrant values on the model's performance. Real-world scenarios demand real-time capabilities for short video popularity prediction [53]. Consequently, the proposed model should exhibit proficiency in promptly processing input data and generating prediction results. Achieving this may necessitate the implementation of optimization techniques, such as parallel computing and model compression, to enhance the operational efficiency of the model. Upholding compliance with privacy regulations and securing explicit user consent are essential considerations when collecting and utilizing user data. Any unauthorized data collection and usage may contravene privacy laws, resulting in potential harm to users [54].

V. CONCLUSION

A. RESEARCH CONTRIBUTION

In the context of cross-cultural communication, this study introduces the innovative CPRP-CNN model, integrating IoT technology and multi-modal data fusion. This model successfully predicts short video popularity with a notable level of accuracy, demonstrating competence in handling intricate tasks. Leveraging multi-layer convolution and pooling operations, the CPRP-CNN model proficiently extracts features from input data, employing fully connected layers for classification, regression, and other tasks [55]. The exploration

of user behavior and social interaction across diverse cultural backgrounds aims to comprehend the global popularity dynamics of short videos. Adopting a multi-modal feature fusion approach involving images, text, and user behavior data, the study captures user preferences in varied cultural contexts and their interaction with short videos. This methodology attains significant success in the realm of popularity prediction [56].

B. FUTURE WORKS AND RESEARCH LIMITATIONS

This study is presently constrained to a specific cultural context and social media platform. Future extensions will broaden the scope to encompass diverse cultures and platforms, facilitating a more exhaustive understanding. Furthermore, the utilization of deep learning models necessitates substantial datasets for effective training. However, certain cultural settings may present challenges in data acquisition, highlighting the significance of addressing issues related to data scarcity. Addressing this challenge represents a valuable avenue for further investigation. In instances of limited data, the judicious selection of representative features pertinent to the target task assumes great importance. Feature selection algorithms may be employed to identify task-relevant features or feature extraction algorithms can be utilized to derive meaningful features from original data.

REFERENCES

[1] K.-M. Giannakopoulou, I. Roussaki, and K. Demestichas, "Internet of Things technologies and machine learning methods for Parkinson's disease diagnosis, monitoring and management: A systematic review," *Sensors*, vol. 22, no. 5, p. 1799, Feb. 2022.

[2] T.-V. Nguyen, N.-N. Dao, V. Dat Tuong, W. Noh, and S. Cho, "User-aware and flexible proactive caching using LSTM and ensemble learning in IoT-MEC networks," *IEEE Internet Things J.*, vol. 9, no. 5, pp. 3251–3269, Mar. 2022.

[3] S. Aminizadeh, A. Heidari, S. Toumaj, M. Darbandi, N. J. Navimipour, M. Rezaei, S. Talebi, P. Azad, and M. Unal, "The applications of machine learning techniques in medical data processing based on distributed computing and the Internet of Things," *Comput. Methods Programs Biomed.*, vol. 241, Nov. 2023, Art. no. 107745.

[4] L. O. Colombo-Mendoza, M. A. Paredes-Valverde, M. D. P. Salas-Zárate, and R. Valencia-García, "Internet of Things-driven data mining for smart crop production prediction in the peasant farming domain," *Appl. Sci.*, vol. 12, no. 4, p. 1940, Feb. 2022.

[5] F. M. Aswad, A. N. Kareem, A. M. Khudhur, B. A. Khalaf, and S. A. Mostafa, "Tree-based machine learning algorithms in the Internet of Things environment for multivariate flood status prediction," *J. Intell. Syst.*, vol. 31, no. 1, pp. 1–14, Nov. 2021.

[6] A. Smerdov, A. Somov, E. Burnaev, B. Zhou, and P. Lukowicz, "Detecting video game player burnout with the use of sensor data and machine learning," *IEEE Internet Things J.*, vol. 8, no. 22, pp. 16680–16691, Nov. 2021.

[7] B. M. Reddy, "Amalgamation of Internet of Things and machine learning for smart healthcare applications—A review," *Int. J Comp. Eng. Sci. Res*, vol. 5, pp. 8–36, Jun. 2023.

[8] J. Liu, J. Bai, H. Li, and B. Sun, "Improved LSTM-based abnormal stream data detection and correction system for Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 18, no. 2, pp. 1282–1290, Feb. 2022.

[9] C. Chen, J. Li, V. Balasubramaniam, Y. Wu, Y. Zhang, and S. Wan, "Contention resolution in Wi-Fi 6-enabled Internet of Things based on deep learning," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5309–5320, Apr. 2021.

[10] T. Han, K. Muhammad, T. Hussain, J. Lloret, and S. W. Baik, "An efficient deep learning framework for intelligent energy management in IoT networks," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3170–3179, Mar. 2021.

[11] S. Latif, M. Driss, W. Boulila, Z. E. Huma, S. S. Jamal, Z. Idrees, and J. Ahmad, "Deep learning for the industrial Internet of Things (IIoT): A comprehensive survey of techniques, implementation frameworks, potential applications, and future directions," *Sensors*, vol. 21, no. 22, p. 7518, Nov. 2021.

[12] B. Chen, L. Liu, M. Sun, and H. Ma, "IoTCache: Toward data-driven network caching for Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10064–10076, Dec. 2019.

[13] V. Chamola, V. Hassija, S. Gupta, A. Goyal, M. Guizani, and B. Sikdar, "Disaster and pandemic management using machine learning: A survey," *IEEE Internet Things J.*, vol. 8, no. 21, pp. 16047–16071, Nov. 2021.

[14] T. J. Saleem and M. A. Chishti, "Deep learning for the Internet of Things: Potential benefits and use-cases," *Digit. Commun. Netw.*, vol. 7, no. 4, pp. 526–542, Nov. 2021.

[15] A. Mellit and S. Kalogirou, "Artificial intelligence and Internet of Things to improve efficacy of diagnosis and remote sensing of solar photovoltaic systems: Challenges, recommendations and future directions," *Renew. Sustain. Energy Rev.*, vol. 143, Jun. 2021, Art. no. 110889.

[16] L. Li, Y. Xu, J. Yin, W. Liang, X. Li, W. Chen, and Z. Han, "Deep reinforcement learning approaches for content caching in cache-enabled D2D networks," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 544–557, Jan. 2020.

[17] Z. Ghaffar, A. Alshahrani, M. Fayaz, A. M. Alghamdi, and J. Gwak, "A topical review on machine learning, software defined networking, Internet of Things applications: Research limitations and challenges," *Electronics*, vol. 10, no. 8, p. 880, Apr. 2021.

[18] I. U. Din, M. Guizani, J. J. P. C. Rodrigues, S. Hassan, and V. V. Korotaev, "Machine learning in the Internet of Things: Designed techniques for smart cities," *Future Gener. Comput. Syst.*, vol. 100, pp. 826–843, Nov. 2019.

[19] M. Waqas, K. Kumar, A. A. Laghari, U. Saeed, M. M. Rind, A. A. Shaikh, F. Hussain, A. Rai, and A. Q. Qazi, "Botnet attack detection in Internet of Things devices over cloud environment via machine learning," *Concurrency Comput., Pract. Exper.*, vol. 34, no. 4, p. e6662, Feb. 2022.

[20] S. M. R. Abidi, Y. Xu, J. Ni, X. Wang, and W. Zhang, "Popularity prediction of movies: From statistical modeling to machine learning techniques," *Multimedia Tools Appl.*, vol. 79, nos. 47–48, pp. 35583–35617, Dec. 2020.

[21] L. Liu, H. Hu, Y. Luo, and Y. Wen, "When wireless video streaming meets AI: A deep learning approach," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 127–133, Apr. 2020.

[22] K. Abbas, M. K. Hasan, A. Abbasi, S. Dong, T. M. Ghazal, S. N. H. S. Abdullah, A. Khan, D. Alboaneen, F. R. A. Ahmed, T. E. Ahmed, and S. Islam, "Co-evolving popularity prediction in temporal bipartite networks: A heuristics based model," *IEEE Access*, vol. 11, pp. 37546–37559, 2023.

[23] K. Brito and P. J. L. Adeodato, "Machine learning for predicting elections in Latin America based on social media engagement and polls," *Government Inf. Quart.*, vol. 40, no. 1, Jan. 2023, Art. no. 101782.

[24] D. Martín-Gutiérrez, G. Hernández Peñaloza, A. Belmonte-Hernández, and F. Álvarez García, "A multimodal end-to-end deep learning architecture for music popularity prediction," *IEEE Access*, vol. 8, pp. 39361–39374, 2020.

[25] A. A. Mubarak, H. Cao, and S. A. M. Ahmed, "Predictive learning analytics using deep learning model in MOOCs' courses videos," *Educ. Inf. Technol.*, vol. 26, no. 1, pp. 371–392, Jan. 2021.

[26] H. Li, S. Cryer, L. Acharya, and J. Raymond, "Video and image classification using atomisation spray image patterns and deep learning," *Biosyst. Eng.*, vol. 200, pp. 13–22, Dec. 2020.

[27] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimedia Tools Appl.*, vol. 80, no. 2, pp. 2887–2905, Jan. 2021.

[28] M. Zhou, G. H. Chen, P. Ferreira, and M. D. Smith, "Consumer behavior in the online classroom: Using video analytics and machine learning to understand the consumption of video courseware," *J. Marketing Res.*, vol. 58, no. 6, pp. 1079–1100, Dec. 2021.

[29] I. de Zarzà, J. de Curtò, and C. T. Calafate, "Socratic video understanding on unmanned aerial vehicles," *Proc. Comput. Sci.*, vol. 225, pp. 144–154, Oct. 2023.

[30] S. A. Hicks, J. M. Andersen, O. Witczak, V. Thambawita, P. Halvorsen, H. L. Hammer, T. B. Haugen, and M. A. Riegler, "Machine learning-based analysis of sperm videos and participant data for male fertility prediction," *Sci. Rep.*, vol. 9, no. 1, p. 16770, Nov. 2019.

[31] P.-Y. Wang, C.-T. Chen, J.-W. Su, T.-Y. Wang, and S.-H. Huang, "Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism," *IEEE Access*, vol. 9, pp. 55244–55259, 2021.

[32] S. Das and A. K. Kolya, "Predicting the pandemic: Sentiment evaluation and predictive analysis from large-scale tweets on COVID-19 by deep convolutional neural network," *Evol. Intell.*, vol. 15, no. 3, pp. 1913–1934, Sep. 2022.

[33] Z. He, T. Shi, J. Xuan, and T. Li, "Research on tool wear prediction based on temperature signals and deep learning," *Wear*, vols. 478–479, Aug. 2021, Art. no. 203902.

[34] N. Barman, E. Jammeh, S. A. Ghorashi, and M. G. Martini, "No-reference video quality estimation based on machine learning for passive gaming video streaming applications," *IEEE Access*, vol. 7, pp. 74511–74527, 2019.

[35] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Trans. Image Process.*, vol. 30, pp. 4449–4464, 2021.

[36] P. Washington, E. Leblanc, K. Dunlap, Y. Penev, A. Kline, K. Paskov, M. W. Sun, B. Chrisman, N. Stockham, M. Varma, C. Voss, N. Haber, and D. P. Wall, "Precision telemedicine through crowdsourced machine learning: Testing variability of crowd workers for video-based autism feature recognition," *J. Personalized Med.*, vol. 10, no. 3, p. 86, Aug. 2020.

[37] P. Mehta, S. Pandya, and K. Kotecha, "Harvesting social media sentiment analysis to enhance stock market prediction using deep learning," *PeerJ Comput. Sci.*, vol. 7, p. e476, Apr. 2021.

[38] M. Shokrolah Shirazi and B. T. Morris, "Trajectory prediction of vehicles turning at intersections using deep neural networks," *Mach. Vis. Appl.*, vol. 30, no. 6, pp. 1097–1109, Sep. 2019.

[39] P. Redhu and K. Kumar, "Short-term traffic flow prediction based on optimized deep learning neural network: PSO-Bi-LSTM," *Phys. A, Stat. Mech. Appl.*, vol. 625, Sep. 2023, Art. no. 129001.

[40] B. K. Yousafzai, M. Hayat, and S. Afzal, "Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student," *Educ. Inf. Technol.*, vol. 25, no. 6, pp. 4677–4697, Nov. 2020.

[41] G. Chen, Q. Kong, N. Xu, and W. Mao, "NPP: A neural popularity prediction model for social media content," *Neurocomputing*, vol. 333, pp. 221–230, Mar. 2019.

[42] D. Maulud and A. M. Abdulazeez, "A review on linear regression comprehensive in machine learning," *J. Appl. Sci. Technol. Trends*, vol. 1, no. 2, pp. 140–147, Dec. 2020.

[43] S. Aggarwal, S. Saluja, V. Gambhir, S. Gupta, and S. P. S. Satia, "Predicting likelihood of psychological disorders in playerunknown's battlegrounds (PUBG) players from Asian countries using supervised machine learning," *Addictive Behaviors*, vol. 101, Feb. 2020, Art. no. 106132.

[44] F. Yang, D. Wang, F. Xu, Z. Huang, and K.-L. Tsui, "Lifespan prediction of lithium-ion batteries based on various extracted features and gradient boosting regression tree model," *J. Power Sources*, vol. 476, Nov. 2020, Art. no. 228654.

[45] A. Stoll and P. Benner, "Machine learning for material characterization with an application for predicting mechanical properties," *GAMM-Mitteilungen*, vol. 44, no. 1, Mar. 2021, Art. no. e202100003.

[46] J. de Curtò, I. de Zarzà, G. Roig, and C. T. Calafate, "Summarization of videos with the signature transform," *Electronics*, vol. 12, no. 7, p. 1735, Apr. 2023.

[47] W. Zhu, Y. Huang, X. Xie, W. Liu, J. Deng, D. Zhang, Z. Wang, and J. Liu, "AutoShot: A short video dataset and state-of-the-art shot boundary detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 2237–2246.

[48] Q. Cai, Z. Xue, C. Zhang, W. Xue, S. Liu, R. Zhan, X. Wang, T. Zuo, W. Xie, D. Zheng, P. Jiang, and K. Gai, "Two-stage constrained actor-critic for short video recommendation," in *Proc. ACM Web Conf.*, Apr. 2023, pp. 865–875.

[49] Q. Ji, S. Zhang, Q. Duan, Y. Gong, Y. Li, X. Xie, J. Bai, C. Huang, and X. Zhao, "Short- and medium-term power demand forecasting with multiple factors based on multi-model fusion," *Mathematics*, vol. 10, no. 12, p. 2148, Jun. 2022.

[50] N. T. P. Hartono, J. Thapa, A. Tiihonen, F. Oviedo, C. Batali, J. J. Yoo, Z. Liu, R. Li, D. F. Marrón, M. G. Bawendi, T. Buonassisi, and S. Sun, "How machine learning can help select capping layers to suppress perovskite degradation," *Nature Commun.*, vol. 11, no. 1, p. 4172, Aug. 2020.

[51] Y. Fang, Z. Li, J. Yan, X. Sui, and H. Liu, "Study of spatio-temporal modeling in video quality assessment," *IEEE Trans. Image Process.*, vol. 32, pp. 2693–2702, 2023.

[52] S. Wang, S. Yang, H. Su, C. Zhao, C. Xu, F. Qian, N. Wang, and Z. Xu, "Robust saliency-driven quality adaptation for mobile 360-degree video streaming," *IEEE Trans. Mobile Comput.*, vol. 23, no. 2, pp. 1312–1329, Feb. 2024.

[53] S. Ye and T. Zhao, "Team knowledge management: How leaders' expertise recognition influences expertise utilization," *Manage. Decis.*, vol. 61, no. 1, pp. 77–96, Jan. 2023.

[54] S. Ye, K. Yao, and J. Xue, "Leveraging empowering leadership to improve employees' improvisational behavior: The role of promotion focus and willingness to take risks," *Psychol. Rep.*, vol. 32, Apr. 2023, Art. no. 003329412311727.

[55] Z. Z. Zhong and E. Y. Zhao, "Collaborative driving mode of sustainable marketing and supply chain management supported by metaverse technology," *IEEE Trans. Eng. Manag.*, vol. 71, pp. 1642–1654, 2024, doi: 10.1109/tem.2023.3337346.

[56] L. Yuan, H. Li, S. Fu, and Z. Zhang, "Learning behavior evaluation model and teaching strategy innovation by social media network following learning psychology," *Frontiers Psychol.*, vol. 13, Jul. 2022, doi: 10.3389/fpsyg.2022.843428.