

Article

DanceTrend: An Integration Framework of Video-Based Body Action Recognition and Color Space Features for Dance Popularity Prediction

Shiyong Ding¹, Xingyu Hou¹, Yujia Liu¹, Wenxuan Zhu², Dong Fang¹, Yusi Fan¹, Kewei Li^{1,*}, Lan Huang¹ and Fengfeng Zhou^{1,3,*} 

- ¹ College of Computer Science and Technology, and Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China; dingsy21@mails.jlu.edu.cn (S.D.); houxy2120@mails.jlu.edu.cn (X.H.); 223040011@link.cuhk.edu.cn (D.F.); fan_yusi@163.com (Y.F.); huanglan@jlu.edu.cn (L.H.)
- ² School of Computer Science and Engineering, and Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications of Ministry of Education, Southeast University, Nanjing 210096, China; zhuwx@seu.edu.cn
- ³ School of Biology and Engineering, Guizhou Medical University, Guiyang 550025, China
- * Correspondence: kwbb1997@gmail.com (K.L.); fengfengzhou@gmail.com or ffzhou@jlu.edu.cn (F.Z.); Tel./Fax: +86-431-8516-6024 (F.Z.)

Abstract: Background: With the rise of user-generated content (UGC) platforms, we are witnessing an unprecedented surge in data. Among various content types, dance videos have emerged as a potent medium for artistic and emotional expression in the Web 2.0 era. Such videos have increasingly become a significant means for users to captivate audiences and amplify their online influence. Given this, predicting the popularity of dance videos on UGC platforms has drawn significant attention. **Methods:** This study postulates that body movement features play a pivotal role in determining the future popularity of dance videos. To test this hypothesis, we design a robust prediction framework DanceTrend to integrate the body movement features with color space information for dance popularity prediction. We utilize the jazz dance videos from the comprehensive AIST++ street dance dataset and segment each dance routine video into individual movements. AlphaPose was chosen as the human posture detection algorithm to help us obtain human motion features from the videos. Then, the ST-GCN (Spatial Temporal Graph Convolutional Network) is harnessed to train the movement classification models. These pre-trained ST-GCN models are applied to extract body movement features from our curated Bilibili dance video dataset. Alongside these body movement features, we integrate color space attributes and user metadata for the final dance popularity prediction task. **Results:** The experimental results endorse our initial hypothesis that the body movement features significantly influence the future popularity of dance videos. A comprehensive evaluation of various feature fusion strategies and diverse classifiers discern that a pre-post fusion hybrid strategy coupled with the XGBoost classifier yields the most optimal outcomes for our dataset.

Keywords: popularity prediction; human action recognition; dance video; feature fusion; XGBoost



Citation: Ding, S.; Hou, X.; Liu, Y.; Zhu, W.; Fang, D.; Fan, Y.; Li, K.; Huang, L.; Zhou, F. DanceTrend: An Integration Framework of Video-Based Body Action Recognition and Color Space Features for Dance Popularity Prediction. *Electronics* **2023**, *12*, 4696. <https://doi.org/10.3390/electronics12224696>

Academic Editor: George A. Tsihrintzis

Received: 20 October 2023

Revised: 12 November 2023

Accepted: 16 November 2023

Published: 18 November 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the Web 2.0 era, rapid advancements in information technology have propelled mobile clients to unprecedented heights [1]. Modern social media platforms are characterized by user-driven content creation and interactions and have overshadowed traditional media as the primary means for information exchange. This transition is primarily attributed to their reduced creative costs and enhanced dissemination capabilities [2]. The transparent nature of user interactions (like forwards and likes) on these platforms under the Web 2.0 framework offers invaluable information for predicting the popularity of social media

content [3]. Among the diverse offerings of user-generated content (UGC) platforms, dance videos have notably risen to mainstream prominence [4].

Dance has been an ancient form of expression, predating the advent of language, imagery, and other artistic manifestations. Its intrinsic capacity to vividly convey profound human emotions renders it unique compared to other art forms [5]. The ever-evolving media landscape has amplified the reach and impact of dance, and the proliferation of variety shows has broadened the demographic appeal of street dance. This genre has now become a popular category of user uploads across media platforms, attracting substantial viewership and engagement [6]. In fact, dance-related topics have attracted researchers' attention for a long time. For example, Costas Panagiotakis et al. [7] proposed a framework for generating beat synchronization dance animation based on visual and audio data analysis in 2013. They identified dancers' joints in videos and matched them with rhythm pulses estimated by the beat-tracking algorithm to generate dance animations of virtual characters.

Video popularity is closely associated with content relevance and user dynamics. While content is the key to information dissemination and determines its spread, user-related metrics such as activity levels and influence also bear considerable impacts on a video's traction [3]. Tsur and Rappoport showed that content resonating with contemporary events or trending topics attracts elevated attention [8]. This observation was further supported by another case study using Weibo [9]. Bakshy et al. discerned a positive correlation between Twitter popularity and follower count [10], and Deng et al. found a strong correlation between Weibo's post virality and the poster's activity level [11]. Despite the plethora of studies on popularity prediction, limited attention has been paid to the predictive analysis of dance video popularity based on human body movements.

This study proposes an integration framework DanceTrend to fuse the human body movement features with other information like color space attributes and user metrics to predict the dance video popularity. We formulate this prediction task as binary classifications of four popularity aspects, i.e., views, likes, favorites, and coin contributions, based on median value post ranking. Our primary dataset is sourced from Bilibili, a popular video platform established in 2009, which generously offers the data for academic endeavors.

The main contributions of this study are as follows:

1. A novel PCA-based feature fusion framework DanceTrend was proposed for the dance video popularity prediction task.
2. DanceTrend effectively integrates three main feature groups, i.e., human body movement features, color space features, and user metrics features.
3. The performance of DanceTrend was evaluated on both public and independently collected dance video datasets.

2. Related Work

2.1. Factors Influencing Popularity

Existing research in popularity prediction underscores three main determinants: user-generated content, release time, and user dynamics [12,13]. These elements are also focal points in this study.

User-Generated Content: Content is the cornerstone of information dissemination, serving as a pivotal medium to capture attention. Yang and Counts discerned that the degree to which users were referenced in social media content directly correlated with the content's attraction [14]. Tan et al. further explored the variants in popularity based on different expressions of identical topics on Twitter [15]. Their findings elucidated that content presentation style significantly influenced its virality. A complementary pattern was observed about a marked preference for visual content on Weibo, where posts interspersed with imagery consistently outperformed posts with limited imagery information in terms of popularity [16].

Release Time: The temporal dimension of content release is important in its receptivity. Factors like the immediacy of content release in relation to current events and temporal patterns influence its virality. Platform recommendation algorithms also favor videos

released in specified timeframes. Canneyt et al. found that the time when information was posted on Twitter and Facebook largely determined its initial popularity [17].

User Dynamics: The content creator's personal metrics, such as their visual presence and follower base, carry significant influence over the content's popularity trajectory. Baskshy et al. ascertained that tweets from users with a substantial follower count on Twitter consistently registered higher dissemination and traction compared to their counterparts with a more limited reach [10].

2.2. Human Action Recognition

The methodology for recognizing human actions is central to our body action extraction process in this study. Human motion recognition has always been a great concern in computational graphics. In a paper published in 2018, Abu Zaher Md Faridee et al. [18] proposed a deep and self-evolving feature learning model HappyFeet based on convolutional neural networks (CNNs) to process motion data captured by wearable sensors. HappyFeet effectively resolved the problem that most wearable sensors don't have enough resolution. For video data, there are now mature frameworks for target recognition and human posture recognition.

YOLO (You Only Look Once) takes inspiration from the human ability to instantaneously identify objects within a visual frame [19]. Traditional object detection algorithms like RCNN (Region-CNN) and Fast RCNN segregate detection results into object categories and positions, while YOLO interprets object detection as a regression task. It utilizes a holistic end-to-end network, seamlessly transitioning from the intake of the original image to the pinpointing of object locations and categories. Finally, it delivers the positions, categories, and confidence probabilities of all objects within the image in a singular glance.

Fang et al. proposed the Regional Multi-Person Pose Estimation (RMPE) algorithm with inaccurate human bounding boxes [20]. Further, they improved it as the AlphaPose to estimate whole-body regional multi-person poses [21]. They utilized the top-down methodology to detect individual borders within visual scope before identifying human poses within each boundary. Enhancements like the symmetric space transformation network (SSTN) and the pose-guided proposals generator (PGPG) for sample amplification further refine its functionality. The initial version of AlphaPose did not work well on multi-target detection. As AlphaPose became a more general human posture recognition project, it was updated with support for multi-target detection methods. The default target detection component of the latest version of AlphaPose is YOLO v3. Fang et al. compared a variety of human posture recognition methods and classified them according to whether the bottom-up method or top-down method is used in posture recognition [21]. Finally, the optimal recognition results of the whole body, foot, face, and hands all appear in the top-down attitude recognition method. These top-down recognition methods use YOLO v3 as the multi-person target recognition method.

The intrinsic nature of the human skeleton has often been represented via graph structures, spurring interest in leveraging graph convolutional neural network (GCN) for human pose detection. The ST-GCN (spatial-temporal GCN) [22] was proposed as an intuitive choice for gesture recognition. It was initially conceptualized for traffic prediction, and empirical studies have showcased the ST-GCN's power in effectively capturing intricate spatio-temporal correlations of human poses [23].

The performance of the human motion recognition framework needs to be evaluated based on the real key point information through human motion capture. Such experiments have also attracted the attention of researchers. In 2020, Rollyn T. Labuguen et al. [24] investigated the performance of the human pose recognition framework OpenPose, comparing the output joint position estimated by OpenPose with the mark-based motion capture data recorded on popular dance movements. Their comparison results show that the average absolute error for each key point is less than 700 mm.

2.3. Multi-Modal Feature Fusion

The fusion of diverse features to predict video popularity has become an increasingly critical area of focus. The general approach in multi-modal feature fusion first involves the extraction and construction of distinct types of features. These are then modeled to establish their correlations with the target label “video popularity” in this study. Such a setup allows for the full utilization of multifaceted information from various modalities. Video visual feature extraction predominantly employs either deep learning-based methods or artificial construction techniques [25,26]. Prominent deep learning architectures such as ResNet-50 [27] and ResNet-101 [28] are frequently deployed for extracting image-based visual features. The obtained high-dimensional feature space may be refined by strategies like retaining the last layer of the CNN encoder as the extracted feature set [29] or applying principal component analysis (PCA) for dimensionality reduction [30,31].

Feature fusion methods can be categorized into two main types: early and late fusion. Early fusion entails the consolidation of multiple high-dimensional features into a reduced-dimensional space. Conversely, late fusion is also known as decision fusion, which focuses on merging the prediction results after a sample has been classified by discriminant classifiers trained on the individual feature groups. The currently popular trend is a hybrid approach that combines the advantages of both early and late fusion [32]. The development of deep learning technologies has been fully integrated with these early, late, and hybrid fusion strategies within the realm of multi-modal feature-based investigations.

2.4. Popularity Prediction Methods

The image/video popularity prediction task can be approached as either a regression or classification problem, depending on the specific objectives and requirements of a given study. Some works have formulated it as a regression problem. Khosla et al. calculated the normalized view count of images as the target labels by fusing the color, gradient, and deep learning features [33]. Gelli et al. fused the visual sentiment and object features to predict the log-transformed number of views on Flickr, and an off-the-shelf support vector regression (SVR) model achieved satisfying performance [34].

Many studies have investigated image/video popularity prediction as a classification problem. Totti et al. simplified the task into a binary classification problem and utilized a random forest classifier to predict the popularity of an image based on its share count on Pinterest [35]. An ensemble framework based on a variety of base classifiers was proposed for the three-tier classification of video playback volume [36]. Jeon et al. separated videos into different classes based on their popularity and release status [37] and trained classification models for both published and newly released videos with XGBoost and deep neural network classifiers, respectively. The classifier XGBoost was also successfully applied to the binary classification of YouTube video popularity after feature selection and fusion [38]. Sarkar et al. introduced a deep neural network framework called ViViD to handle the multi-modal features and perform multiclass prediction for video popularity [39].

3. Materials and Methods

This study formulated the popularity prediction task of dance videos as a binary classification problem. There are four popularity indicators, i.e., views, likes, favorites, and coin contributions, for each video. The samples are sorted by the ascendent order of each popularity indicator, and the threshold is established at their median value. Videos with an index above this threshold are considered as “popular”, while those below it are classified as unpopular.

3.1. Datasets

This study trained and evaluated the proposed framework DanceTrend using two datasets, i.e., AIST++ and Bilibili Dance Video (BDV).

The Google team released the AIST++ dataset [40], the most comprehensive dataset of street dance videos to date. The dataset encompasses a wide array of hip-hop dance styles,

including but not limited to Breaking, Popping, Locking, Hip-hop, House, Waacking, and Jazz. The dataset comprises 1.1 million frames and totals approximately 5 h in duration. The 1408 sequences of dance videos span basic to advanced levels of dance choreography. AIST++ has primarily been used in motion generation research, serving purposes such as movement transfer, dance classification, and movement classification. This study represents the first instance where the dataset is segmented and tailored specifically for training motion classification models.

This study constructed the second dataset of Bilibili dance videos (BDV) from the Bilibili platform. This dataset comprises 769 dance videos carefully gathered between 1 February 2023 and 27 February 2023 from the Bilibili platform's "Jazz" category. Each video in this dataset was selected based on its recency, clear recording quality, and full-body visibility. Popularity data for these videos were collected 31 days post-release, ensuring a reasonable timeframe for gauging their reception.

The Bilibili dataset is feature-rich, and each sample is characterized by 54 different features. These features are categorized into three main domains: action-related features, visual (color) features, and user-specific information. These comprehensive features serve as the basis for our exploratory analysis and further model training.

3.2. Performance Metrics

This study primarily focuses on Accuracy (ACC) and F1 score as the key evaluation metrics for binary classification. These metrics provide an overall view of how well a binary classification model performs in correctly identifying both positive and negative samples. Below are detailed explanations and mathematical formulations for the metrics.

Accuracy is a straightforward measure that quantifies the overall effectiveness of the model in making correct predictions for both positive and negative samples. It is calculated using the formula: Accuracy (ACC) = (TP + TN)/(TP + FP + TN + FN), where TP, FN, TN, and FP are the numbers of true positives, false negatives, true negatives, and false positives, respectively.

The F1 score is a balanced metric that computes the harmonic mean of Precision and Recall. It ranges between 0 and 1, with higher values indicating superior performance. The metrics Precision and Recall are defined as Precision = TP/(TP + FP) and Recall = TP/(TP + FN). Then, we can calculate F1 = 2 × Precision × Recall/(Precision + Recall).

In this study, PCA is used as the feature fusion method, and its calculation formula is as follows.

First, we need to standardize the dataset. Assuming the number of samples is n and the number of features is m , we can obtain a data matrix of size $n \times m$. We need to calculate the average value by column, and we obtain the column mean value $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$. We also need to obtain the column-specific standard deviation $S_j = \sqrt{\frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n-1}}$. Then, we can obtain the standardized data $X_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j}$.

Second, we calculate the covariance matrix of the standardized sample matrix R with size $m \times m$, in which $r_{ij} = \frac{1}{n-1} \sum_{k=1}^n X_{ki} X_{kj}$. After we determine R , we can calculate its eigenvalues λ_1 to λ_m and eigenvectors a_1 to a_m .

Finally, we calculate the contribution degree $\frac{\lambda_i}{\sum_{k=1}^m \lambda_k}$ of each eigenvalue and construct the principal components by taking the eigenvector with the top N contribution degrees. The i th principal component = $a_{1i}X_1 + a_{2i}X_2 + \dots + a_{mi}X_m$ ($i \leq m$), and N is a manually set parameter less than or equal to m .

The Maximum Information Coefficient (MIC) is used in the feature selection step. We first define the term mutual information, which describes the amount of information in the corresponding part of the two subsystems of the same system. The calculation of mutual information is defined as follows:

$$I(x; y) = \int p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} dx dy$$

where $p(x, y)$ is the joint probability between the variables x and y . Next, the calculation formula of MIC is given:

$$mic(x; y) = \max_{a*b < B} \frac{I(x; y)}{\log_2 \min(a, b)},$$

where a and b are the numbers of cells in the x and y directions divided when the two variables are discretized into a two-dimensional space and represented by a scatter plot, representing the grid distribution; B is a variable whose size is about 0.6 power of the amount of data.

3.3. Sample Construction

The construction of appropriate video samples is crucial for the accuracy and reliability of the prediction models. The beginning parts of the videos are typically added by the publishers during the video release and do not necessarily pertain to the content of the dance itself. This study considers the beginning parts of videos redundant and systematically removes them from the dataset. The following editing rules are applied based on the duration of each video:

Duration \leq 25 s: No trimming of the beginning part is performed. Videos of this length usually consist only of a dance clip, and as such, they are used in their entirety.

25 s < Duration \leq 30 s: The first 5 s is trimmed. This is based on the assumption that videos longer than 25 s are likely to start with a redundant prolog.

30 s < Duration \leq 60 s: The first 10 s is removed. Videos within this duration are often more formally constructed, and therefore, longer introductory prologs are expected.

Duration > 60 s: The initial 15 s is cut. For videos exceeding one minute, it is generally assumed that the prologue will be even more elaborate, justifying the longer cut.

By applying these rules systematically, we aim to ensure that the videos used in our training dataset focus solely on the relevant features of dance movements, thereby enhancing the quality and reliability of our machine-learning models.

3.4. Human Body Movement Features

Monika Wysoczanska and Tomasz Trzcinski extracted human skeleton features from dance video samples as video features in multi-modal features. Their data showed that the human movements represented by human skeleton features in the video frame contributed useful information for the downstream prediction tasks [41]. In 2022, Davide Moltisanti et al. released the Breaking competition data set for dance movement synthesis using the Red Bull BC ONE competition, in which they used automatic and manual annotations to obtain the key features of the dancers' human bones in the video [42]. This facilitated more consistent and accurate performance for dance movement synthesis. The data showed that the key points of the human skeleton were extremely useful for the recognition of dance movements.

The accurate extraction of features related to human body movement is pivotal to our study. We partitioned the jazz dance videos from the AIST++ dataset into smaller clips, each corresponding to a distinct dance movement. These clips were then manually assessed to categorize the involvement of specific body parts in each movement. Specifically, we concentrated on the movements of five primary body regions: the head, chest, arms, pelvic area, and legs.

For instance, after capturing a specific dance movement in a video clip, we evaluate whether the movement involves head motion. If it does, the clip is added to a dataset of positive samples for head action. If not, it is placed in a dataset of negative samples. Similar datasets were created for the other four body regions: chest, arms, pelvic area, and legs. The detailed information of the datasets is shown in Table 1.

Table 1. Details of each body region. The table contains the specific content of videos and the number of sample videos.

Body Region	Positive Samples		Negative Samples	
	Content	Counts	Content	Counts
Head	Videos of actions with head movement	48	Videos of actions without head movement	49
Chest	Videos of actions with chest movement	59	Videos of actions without chest movement	44
Arms	Videos of actions with arms movement	68	Videos of actions without arms movement	37
Pelvic area	Videos of actions with pelvic area movement	47	Videos of actions without pelvic area movement	56
Legs	Videos of actions with leg movement	59	Videos of actions without leg movement	43

We employed a spatial temporal graph convolutional network (ST-GCN) to train individual classification networks for each body part, using the curated sets of positive and negative samples. These ST-GCN models aim to classify whether a given body part is engaged in movement within a query clip.

Each video sample is divided into three equally sized segments. The movement of each of the five body parts is assessed at intervals of 12 frames within each segment. Subsequently, we calculate the proportion of time each body part is in motion within each segment. This yields a total of 15 action features, encompassing the time-based activity of the five main body regions across the three segments of each video clip.

By employing this rigorous methodology, we aim to ensure a comprehensive understanding of body movements and to improve the performance and explainability of the proposed DanceTrend framework for the dance video popularity prediction task.

3.5. Color Space Features

Besides the human body movement features described above, Monika Wysoczanska and Tomasz Trzcinski also extracted RGB channel features from dance video samples to represent contextual information [41]. Cheng et al. investigated the micro-video popularity prediction task by extracting the features of three modalities, i.e., audio, text, and video. For video images, they extracted the features by grouping the colors into 50 unique tones in a single RGB channel [43]. Kuo et al. chose to transform RGB features into HSV features to predict the coloring images. They showed that both RGB features and HSV features expressed color information from different dimensions [44].

Color composition in dance videos can significantly impact the viewer's perception and, consequently, the video's popularity. Our analysis incorporates a variety of color space features extracted from the video frames, specifically focusing on grayscale, RGB (Red, Green, Blue), and HSV (Hue, Saturation, Value) color spaces.

Grayscale characteristics offer a foundational understanding of a video's color composition. For each video frame, we calculate both the average and maximum grayscale pixel values. Additionally, we designate pixels with grayscale values not larger than 40 as "dark pixels". For each frame, we calculate and retain the proportion of such dark pixels to the total number of pixels. Finally, we generate four features: the average and maximum proportion of dark pixels across all frames, along with the average and maximum grayscale values.

The RGB color space is based on the three primary colors Red/Green/Blue and is a commonly employed model for digital media. We extract and calculate the average and maximum values of each RGB component for each video frame. After the analysis of the entire video sample, six features are constructed, representing the average and maximum values of the Red/Green/Blue components across all frames.

The HSV color space offers another representation of color, encoding Hue, Saturation, and Value (or brightness). The HSV color space divides hues into degrees, ranging from 0 to 360. For the purpose of this study, we segment the hue spectrum into six equal units, each spanning 60 degrees. Within each unit, we extract the average and maximum values for both Saturation (S) and Value (V) for each frame. This leads to four features per hue unit, i.e., average and maximum S, and average and maximum V. A total of 24 HSV features can be obtained for each sample.

3.6. User Metrics and Additional Features

To offer a more comprehensive analysis of dance video popularity, this study extends beyond the realm of video content to consider various user- and context-specific features. These features encompass the uploader's fan base size, the gender of individuals featured in the video, the video's duration, and a metric to gauge the visual appeal of the performers.

Fan Base Size: The size of the uploader's fan base can significantly influence a video's popularity. We capture this count by recording the number of fans that the uploader has at the time of the video's release.

Gender of Performers: The performers' genders can also be a crucial factor in determining video appeal. Therefore, we classify the genders of the characters in each video into one of the three categories: male, female, or both.

Video Duration: Video length, measured in seconds, is another variable that has a potential impact on viewer engagement and the video's overall popularity.

Visual Appeal Metric: The performers' visual appeal is anticipated to have a potential impact on a video's popularity. We employ Baidu's facial attractiveness scoring API. We randomly select 10 frames from each video and apply the API to obtain the attractiveness scores for the individual features. Both the average and the maximum attractiveness scores across these frames are calculated as the features.

We integrate these user- and context-specific features to provide a multifaceted perspective on the factors that contribute to dance video popularity. This holistic approach enhances the robustness and explainability of our DanceTrend models.

3.7. Feature Fusion Strategy

We adopted a principal component analysis (PCA)-based feature fusion strategy to enrich the following three main groups: human body movement features, color space features, and user metrics features. These features were constructed from the same dataset and could carry inherent connections in such structured data that needed further processing [26,45].

Each feature set in Figure 1 was transformed by PCA, and the same number of principal component (PC) features was obtained. Then, a feature selection and an XGBoost classifier were employed to process each of the three main groups, i.e., human body movement features, color space features, and user metrics features. The weighted result of three predictions was calculated by summing of the three main feature groups: human body movement features, color space features, and user metrics features, with the weights 0.15, 0.20, and 0.65, respectively.

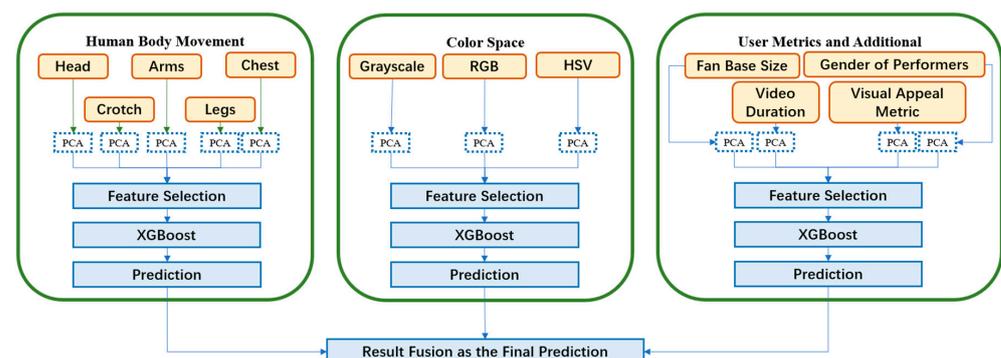


Figure 1. Flowchart representing the feature fusion and classification strategy.

3.8. Experimental Procedure

The complete workflow of this study is illustrated in Figure 2. The experimental design encompasses several stages, from data acquisition to feature extraction, model training, and ultimately to predictive analysis. Below are the key steps in our experimental procedure.

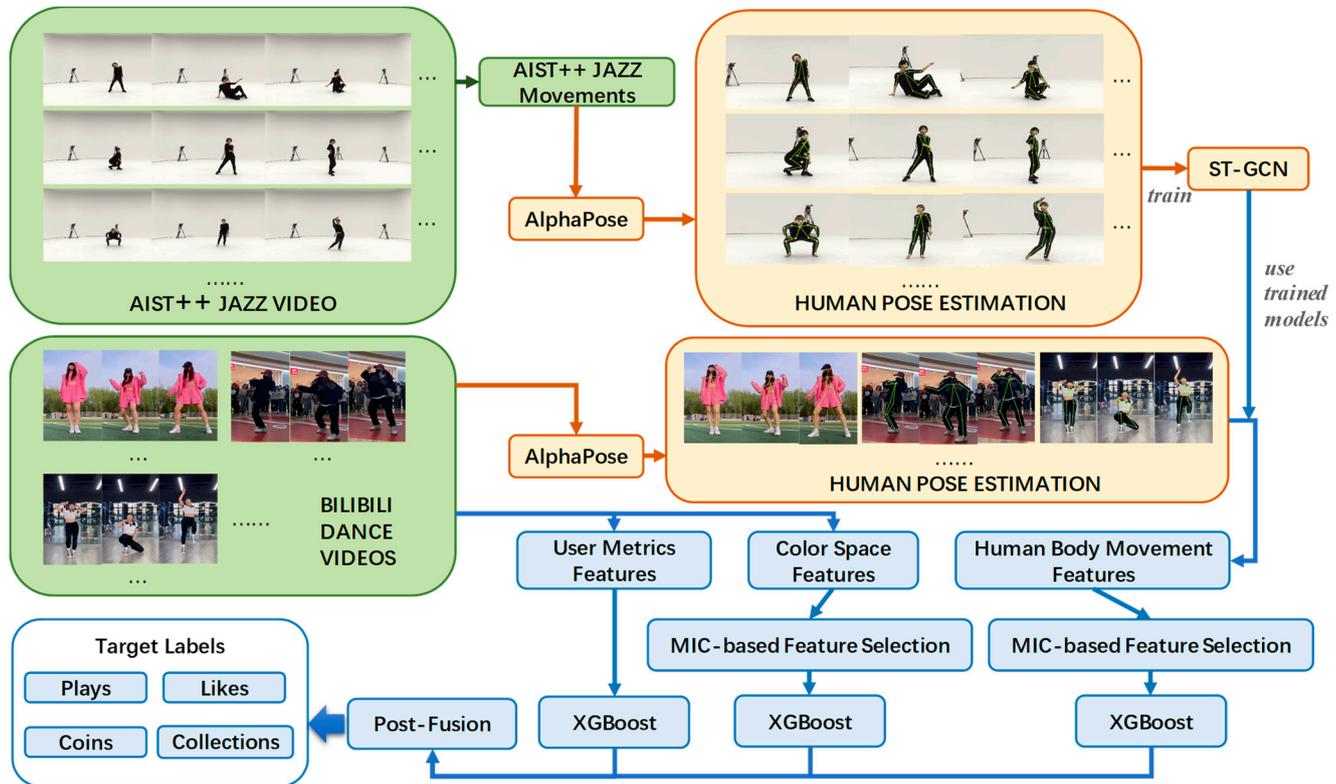


Figure 2. Flowchart outlining the comprehensive experimental design employed in this study. The Bilibili dance videos only take part of the videos that have been authorized for display, in order to protect the personal privacy of all bloggers.

The dance videos from the two datasets AIST++ and BDV were acquired and preprocessed. Initially, jazz dance videos from the AIST++ dataset were segmented into smaller video clips, each corresponding to individual dance movements. For each clip, we labeled movements involving the head, chest, pelvic area, arm, and leg. These annotated clips serve as the training set for the subsequent human body movement classification networks. Both the AIST++ and BDV datasets were processed through a human pose estimation framework, which integrates YOLO for object detection and AlphaPose for keypoint recognition.

The Spatio-Temporal Graph Convolutional Networks (ST-GCNs) were trained using the body keypoint data extracted from the five annotated sets of dance movements in the AIST++ dataset. The trained networks were then employed to evaluate body movements in videos from the BDV dataset. Each BDV video clip was partitioned into three equal-length segments, and a sampling rate of 12 frames per second was used to assess movements across five body parts. Proportional movement metrics for each body part were subsequently calculated.

A comprehensive set of 54 features was collected from the above-generated human body movement features, together with the color space features and the user metrics features. These were then screened by the Maximum Information Coefficient (MIC) [46]. Only 60% of features with the largest MIC values were retained in each of the two groups: human body movement features and color space features. All the user metrics features were preserved.

The refined feature set was used for the three separate predictive models for the three main feature groups. Their prediction results were used by a weighted decision-making

process for the four target labels: number of plays, number of likes, number of favorites, and number of coins.

4. Results and Discussion

4.1. Comparative Experiment

This study employed three groups of features, i.e., human body movement, color space, and user metrics. These features were transformed by PCA before the XGBoost-based predictions. We compared the proposed DanceTrend models with multiple classifiers using the three groups of features without the PCA transformation. The evaluated classifiers included Random Forest (RF), XGBoost, and Decision Tree (DT). Support Vector Machine (SVM) and LightGBM were popularly used in many classification studies [47,48] and were also evaluated in this study. Artificial Neural Network (ANN) was compared for its prediction capacity of the multi-modal data in this study, since ANN has been commonly used for classifying multi-modal data [49].

Figure 3 visualizes the performance metrics across four popularity indicators: likes, coins, favorites, and plays. Our DanceTrend models achieved classification accuracies of 0.8312, 0.7922, 0.8831, and 0.9026 for these respective indicators, outperforming all the comparison models. While the prediction accuracy of DanceTrend for the number of likes was equal to the best-performing comparison model LightGBM, our F1 score marginally exceeded LightGBM by 0.0002. Our model’s prediction accuracy surpassed the best-performing comparison model XGBoost by 0.0390, and the F1 score also exhibited a similar margin of improvement. DanceTrend achieved the same improvement of 0.0714 in ACC than the best-performing comparison models for the numbers of favorites (LightGBM) and plays (XGBoost).

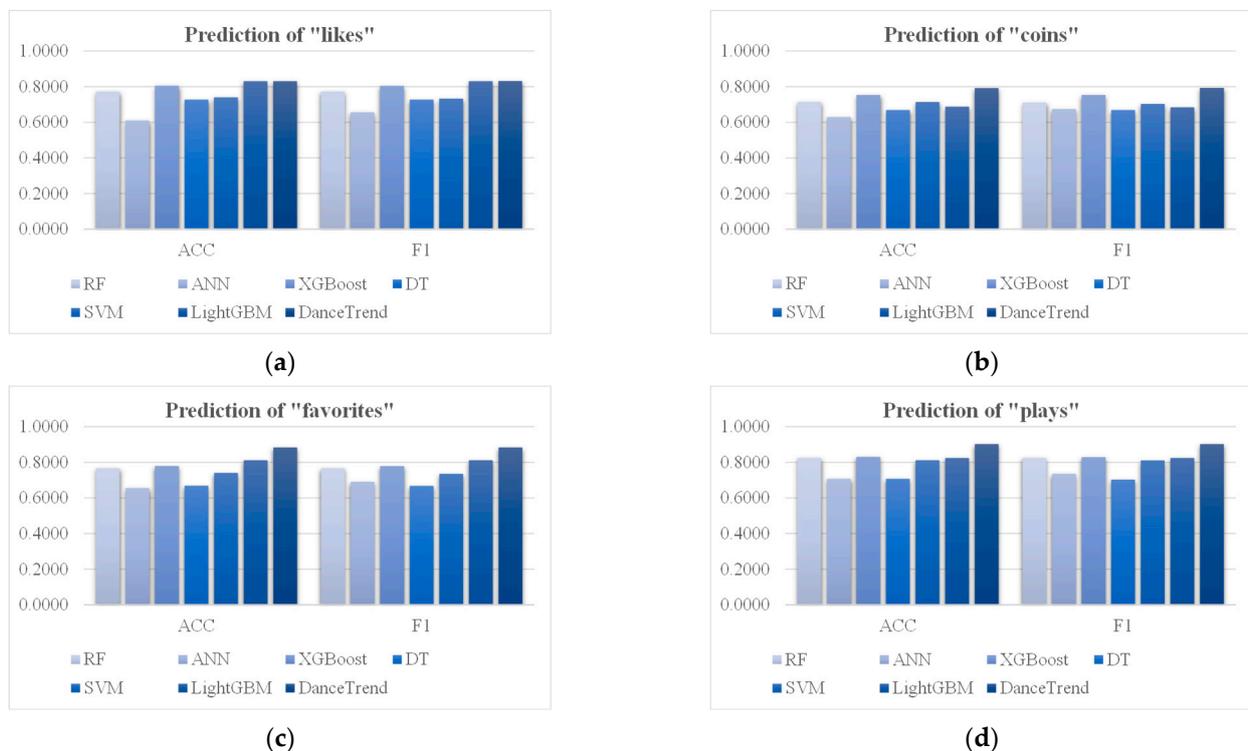


Figure 3. Comparative performance illustrating the efficacy of our approach versus alternative methods across four key indicators of video popularity. The popularity indicators include the numbers of (a) likes, (b) coins, (c) favorites, and (d) plays. The horizontal axis gives the two prediction performance metrics ACC and F1. The vertical axis gives the values of the two metrics.

Taking into account the four popularity indicators collectively, our DanceTrend algorithm demonstrates superior performance metrics and affirms its efficacy in predicting video popularity.

4.2. Determination of Feature Selection Threshold

Maximum Information Coefficient (MIC) can detect both linear and nonlinear correlations between two variables, but it is also sensitive to weak or even false positive corrections [46]. The diversified types of the three main feature groups in this study rely on the MIC's strong capability to detect nonlinear relationships between variables. So, we chose MIC as the indicator of feature selection.

Among these three groups of features, the MIC between the number of fans and predictive indicators is the highest, followed by the video time and the level of people's appearance, which all belong to user metrics features. Among the color features, the S value and V value features divided into red blocks by H value have the highest MIC value. Among the action features, hip and head movements are the top two most significantly correlated with popularity.

We evaluate how feature selection may impact the prediction tasks of the three main feature groups (Figure 4).

There are 15 human body movement features in total. The selection of the top 60% of MIC-ranked features achieved the best ACC values on three video popularity indicators, likes, coins and favorites, while achieving the second-best ACC = 0.5649 on the indicator "plays", which was slightly worse than the best one (ACC = 0.5714). The top 60% of MIC-ranked features equated to retraining nine features.

The dataset initially contained 34 color space features. No singular selection proportion yielded the highest accuracy consistently across all four indicators. The top 60% MIC-ranked color space features consistently ranked within the top three in performance and reached the best average ranks for ACC (rank 2.25) and F1 (rank 2.50) over the four popularity indicators. As such, 20 color space features were ultimately selected.

Given that there were only five user metrics features, the retention ratio led to duplicated outcomes. For instance, 30%, 40%, and 50% retention would all yield two features. But all four subfigures of Figure 4 suggested that the best prediction performance was achieved using all the five user metrics features.

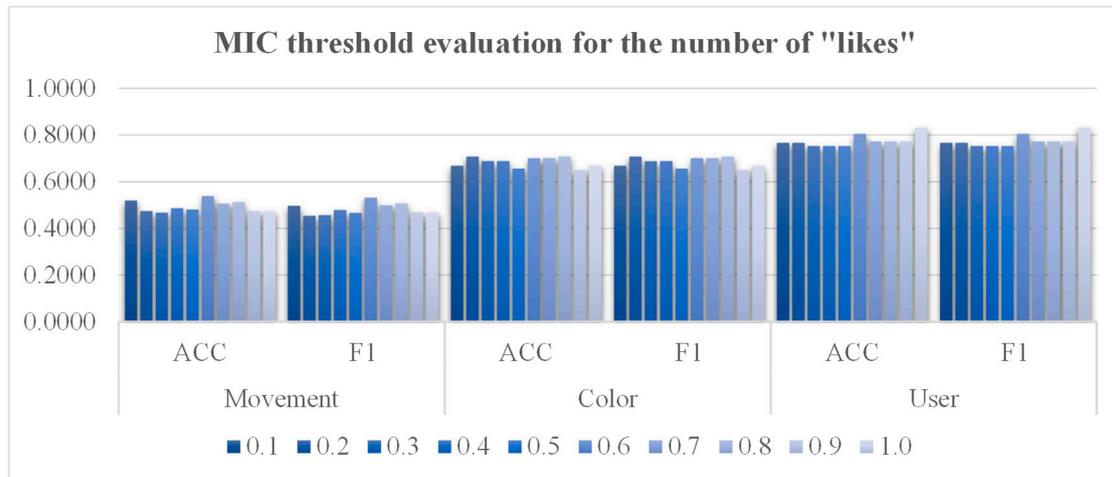
In summary, features within the top 60% MIC values for human body movement and color space features were retained, along with all the user metrics features. This yields a final composite of 34 features for the subsequent analysis.

4.3. Evaluation of Human Body Movement Classifications

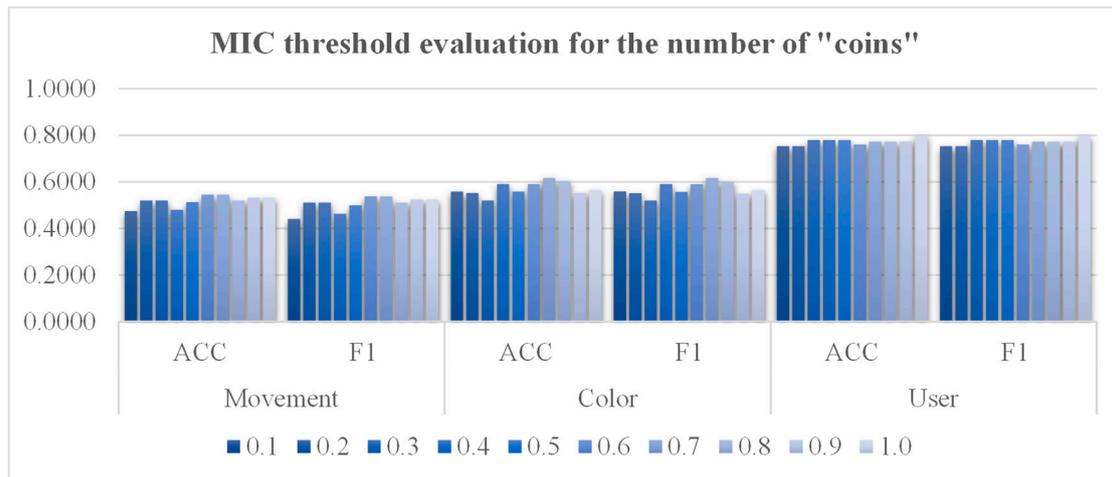
We trained separate classification networks to identify the movements in five body parts: the arms, chest, pelvic area, legs, and head. The networks were trained for 15 epochs, with the results validated using a test set. These outcomes are illustrated in Figure 5.

The network trained for head movement classification shows unstable fluctuations during the first eight epochs (Figure 5e). However, this instability resolves over time, trending toward increased and stable accuracies after the eighth epoch. The variability in the head movement classification could be attributed to its unique challenges. Specifically, the key points on the head are more densely clustered than those on other body parts. Furthermore, the head's smaller volume compared to other body segments may contribute to the increased difficulty in its accurate classification. The movement classification networks trained for the other four body parts quickly converge to the best accuracies on both training and testing subsets. The dataset was partitioned into training and testing subsets at a ratio of 4:1. This distribution results in approximately 20 samples within the testing subset, as outlined in Table 1. It is observed that, during certain epochs, the testing ACC marginally surpasses the training ACC. This phenomenon can be attributed to the relatively small size of the testing subset. Given the limited number of testing samples, even

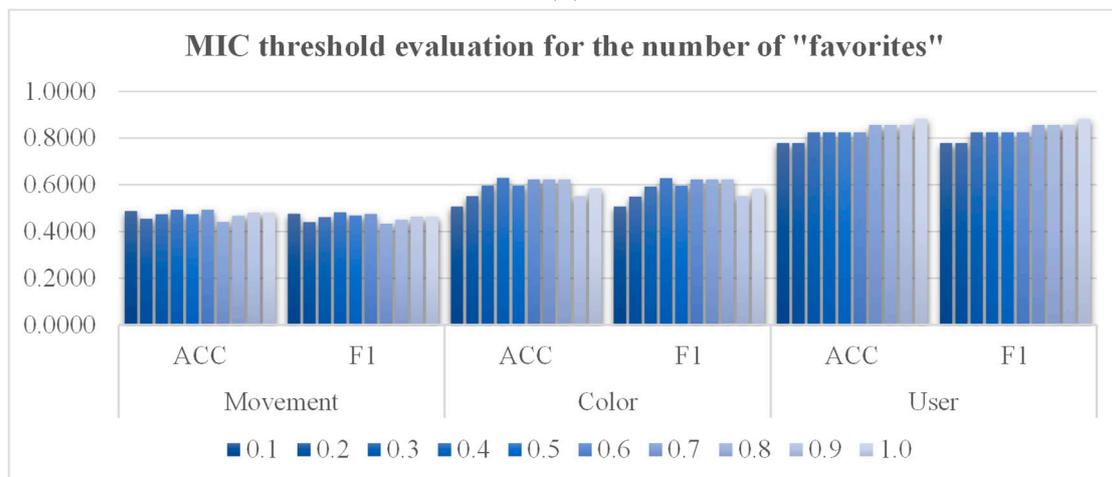
a discrepancy of one or two samples can manifest as a notable variation in the resultant ACC calculation.



(a)

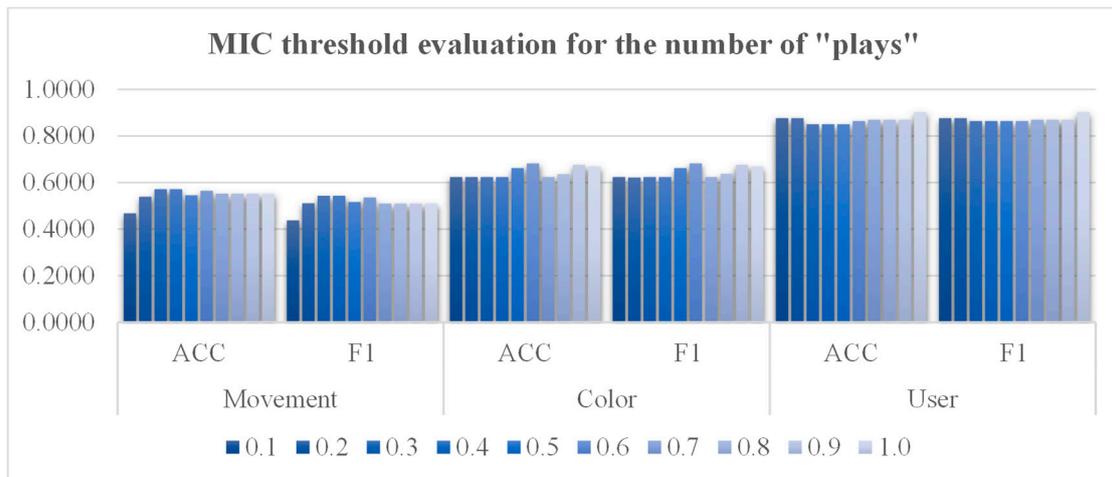


(b)



(c)

Figure 4. Cont.



(d)

Figure 4. Contribution evaluation of feature selection. The horizontal axis indicates the performance metrics ACC and F1 for the three groups of features: human body movements (Movement), color space (Color), and user metrics (User). The vertical axis indicates the values of the two performance metrics ACC and F1. The data series between 0.1 and 1.0 indicates the proportion of the total features in that feature group. If this proportion of the features is smaller than one, we extract one feature for the prediction. The plots are for the four video popularity indicators, i.e., (a) likes, (b) coins, (c) favorites, and (d) plays.

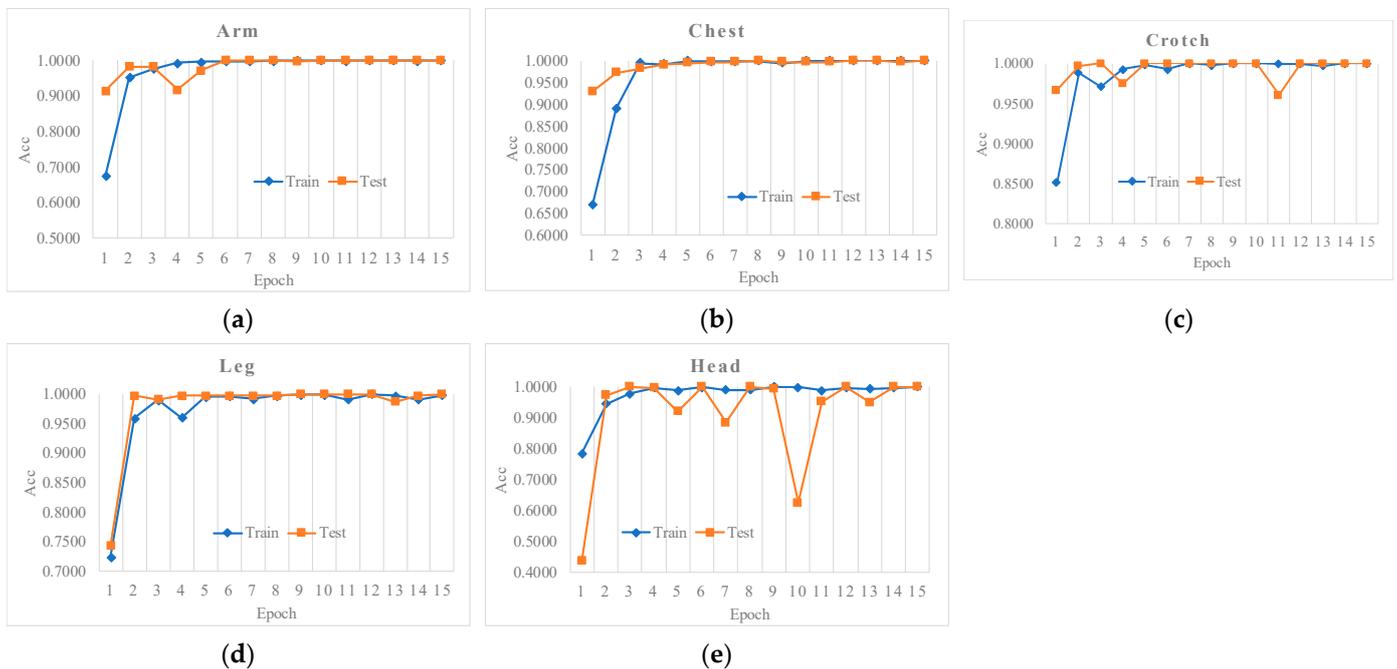


Figure 5. Training performance of the human body movement classification networks. The body parts undergoing binary network training are identified on the left-hand side of the charts. The horizontal axis represents the number of epochs the networks were trained for. The evaluated performance metric is ACC, indicated in the vertical axis. The subfigures correspond to the movement classifications of the (a) arms, (b) chest, (c) pelvic area, (d) legs, and (e) head.

Notwithstanding these epoch-specific fluctuations, there is a consistent enhancement in the testing ACC concurrent with the progression of epochs. This ascending trajectory

aligns with the training ACC trend, indicating robust model generalization despite the dataset's constraints. Therefore, the training outcomes are deemed to be valid and reliable.

4.4. Contribution of Human Body Movement Features

We empirically assess the contribution of incorporating the human body movement features in the video popularity prediction tasks. The above sections showed that the human body movement features achieved worse than the other two feature groups on the video popularity prediction tasks. This section executed an ablation study that specifically focused on the video popularity predictions with and without the human body movement features. The 10-fold cross validation strategy was used to evaluate the convolutional neural network (CNN) model on the predictions across all four popularity indicators.

As illustrated in Figure 6a, not only does the model achieve higher accuracy rates, but it also demonstrates superior performance in F1 score measurements when the human body movement features are integrated. The largest improvement in ACC (0.0558) is achieved for the popularity indicator “plays” when the human body movement features are used.

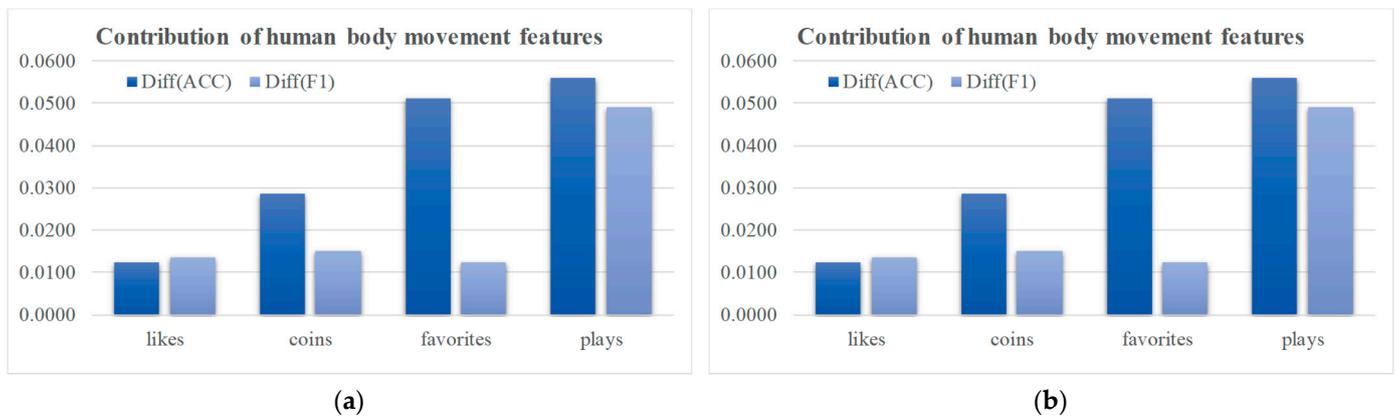


Figure 6. Contribution evaluation of the different feature types to the popularity prediction task. (a) The human body movement features and (b) the color features are evaluated in the popularity prediction task. The horizontal axis represents the four popularity indicators. The vertical axis gives the difference values of the performance metrics ACC and F1 between the baseline CNN model using all three main feature groups and that excluding the evaluated feature group. A positive value indicates that the inclusion of the evaluated feature group achieves better performance than the model without the feature group.

In summary, the human body movement features positively contribute to the video popularity predictions across all the four popularity indicators.

4.5. Contribution of Color Features

We also empirically evaluated the contribution of color features to the video popularity prediction task. This ablation experiment focused on video popularity prediction with and without color features. Similarly, we use a 10-fold cross validation strategy to evaluate the prediction of a convolutional neural network (CNN) model for four popularity indicators.

As shown in Figure 6b, color features also have a positive effect on the prediction of popularity. The two performance metrics ACC and F1 are decreased by at least 0.0214 when the color features are not included in the DanceTrend framework. The largest ACC improvement (0.0409) is achieved in the popularity indicator “plays” and the largest F1 improvement (0.0574) was achieved in the popularity indicator “likes”.

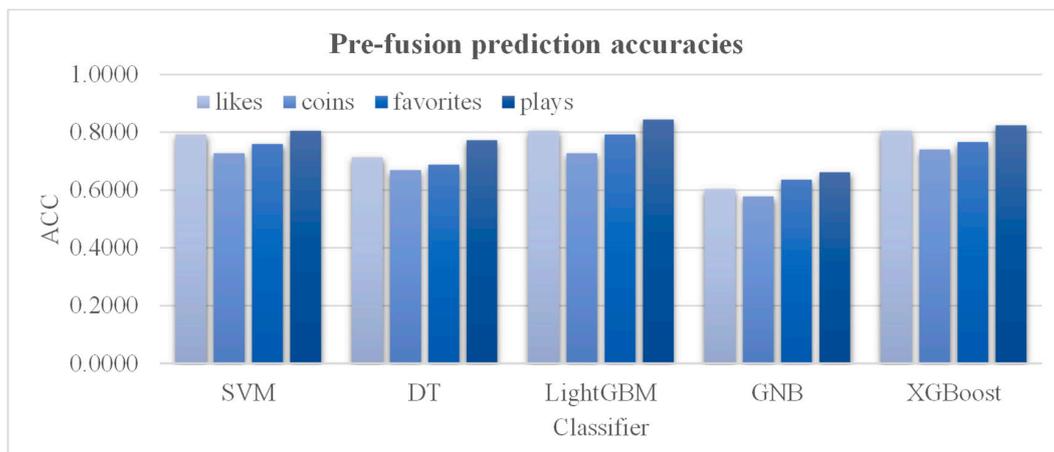
We conclude that color features contribute positively to the dance video popularity predictions for all four popularity metrics.

4.6. Comparative Evaluation of Classifiers across Fusion Strategies

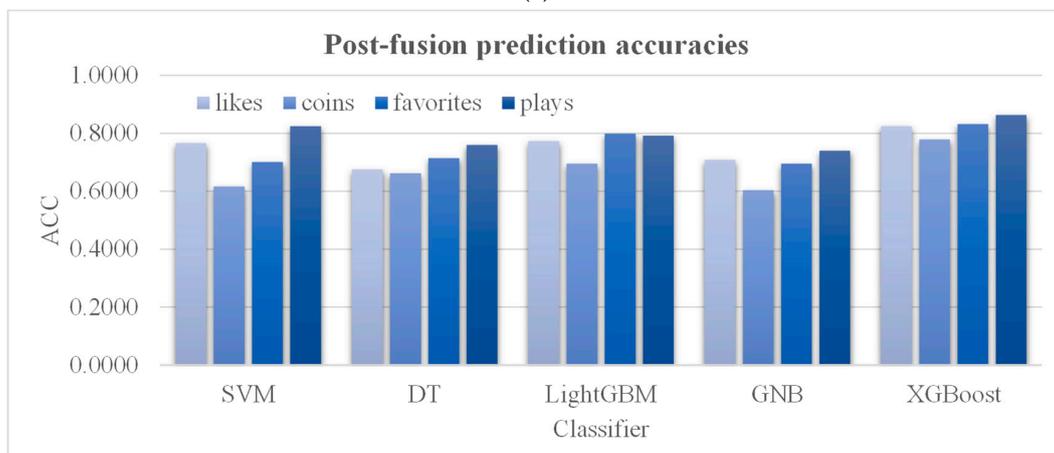
This section evaluated the prediction accuracies of the classifiers when different fusion strategies were employed. We tested five classifiers (including SVM, DT, LightGBM, GNB, and XGBoost) against four popularity indicators: likes, coins, favorites, and plays.

Three fusion strategies were evaluated. The pre-fusion strategy combined the three main feature groups, i.e., human body movement, color space, and user metrics, to build the classification model. The post-fusion strategy involved building a classification model for each of the three feature groups and then generating the final predictions using the weighted sum of the three classification models. This study chose the hybrid fusion strategy, which included transforming the three feature groups separately, building classification models over the three transformed feature groups, and generating the weighted sum of the three predictors.

Figure 7 demonstrates that the different classifiers excelled under various fusion strategies. When the pre-fusion strategy was employed, LightGBM achieved the best accuracies for the three popularity indicators, i.e., likes (ACC = 0.8052), favorites (ACC = 0.7922), and plays (ACC = 0.8442). But it only achieved the second-best ACC (0.7273) for the prediction task for coins. XGBoost yielded the best accuracy, 0.7403, for the prediction task for coins. In the scenarios of both post-fusion and hybrid fusion, XGBoost outperformed the other classifiers across all four popularity indicators. The best classification accuracies of XGBoost were 0.8247, 0.7792, 0.8312, and 0.8636 for the popularity prediction tasks of likes, coins, favorites, and plays. These prediction accuracies were further improved to 0.8312, 0.7922, 0.8831, and 0.9026 under the hybrid fusion strategy in this study.

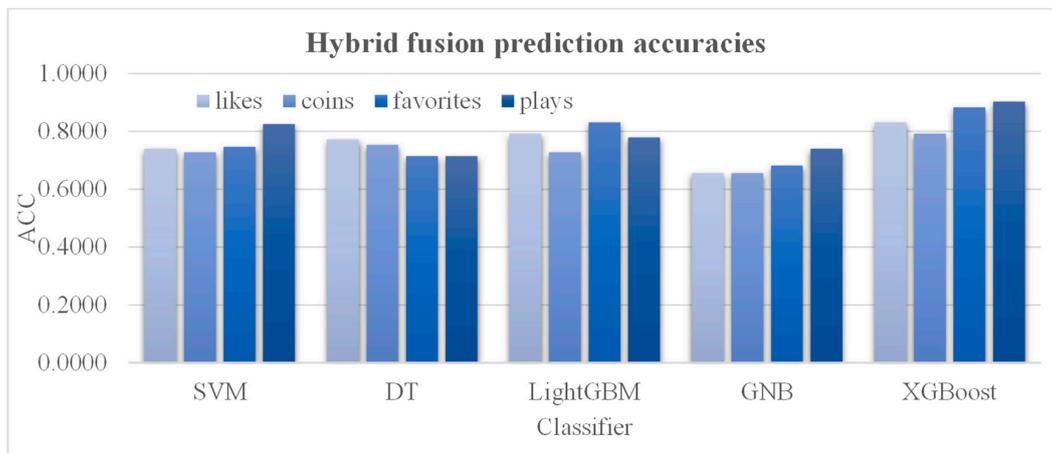


(a)



(b)

Figure 7. Cont.



(c)

Figure 7. Comparing classification accuracies under varying fusion strategies. The histograms illustrate the performances of the evaluated classifiers under the (a) pre-fusion, (b) post-fusion, and (c) hybrid fusion strategies.

XGBoost achieved the overall best classification performance against the other classifiers for the prediction tasks of the four video popularity indicators.

4.7. Determining the Optimal Feature Fusion Method

We also evaluated the impacts of three feature fusion strategies, i.e., pre-fusion only, post-fusion only, and a hybrid approach that combined the ideas of both pre- and post-fusion. These strategies were subsequently evaluated for their performance across four popularity indicators: likes, coins, favorites, and plays.

Figure 8 illustrates the comparative outcomes across the three fusion strategies. The hybrid fusion strategy used in the proposed DanceTrend framework outperformed the pre-fusion and post-fusion methods regarding both accuracy and F1 score across all four popularity indicators.

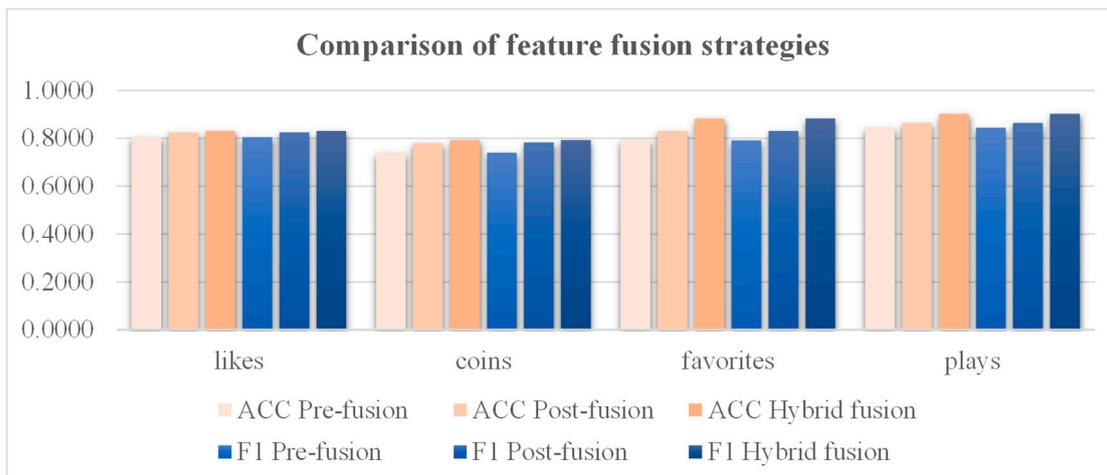


Figure 8. Comparative summary of the accuracy and F1 values achieved by different feature fusion strategies in predicting the four popularity indicators. The horizontal axis gives the four popularity indicators, while the vertical axis gives the values of the accuracies and F1 scores.

The hybrid fusion strategy achieved an accuracy of 0.8312 for the popularity indicator likes, which is 0.0261 higher than the pre-fusion (0.8052) and 0.0065 higher than the post-fusion (0.8247) strategies. For predicting the indicator coins, the hybrid fusion achieved

an accuracy of 0.7922, outperforming the pre-fusion method by 0.0519 (0.7403) and the post-fusion method by 0.0130 (0.7792). Larger improvements were achieved by the hybrid fusion strategy, with an accuracy of 0.8831 for predicting the popularity indicator favorites, which is 0.0909 and 0.0519 higher than the pre-fusion (0.7922) and post-fusion (0.8312) strategies, respectively. The hybrid fusion strategy even achieved a very high accuracy of 0.9026 for predicting the popularity indicator plays, besting the pre-fusion strategy by 0.0584 and the post-fusion strategy by 0.0390. A similar trend was observed for the other prediction performance metric F1 score.

5. Conclusions

This study delves into the topic of dance video popularity prediction, employing human body movement recognition techniques. We showed the positive contribution of the movement features to the prediction of dance video popularity. Additionally, this study investigates the relative merits of feature fusion strategies and classifiers, revealing that our proposed DanceTrend framework significantly outperforms the other approaches, with accuracies of 83.12%, 90.26%, 88.31%, and 79.22% for the four popularity indicators, i.e., “likes”, “plays”, “favorites”, and “coins”, respectively.

While our study proposes a novel framework DanceTrend for the prediction task of dance video popularity, it also comes with its set of limitations. Firstly, this study defines the human body movement features by the movements of only five primary body regions: the head, chest, arms, pelvic area, and legs. The intricate and free-form nature of dance movement adds an inherent layer of complexity to the feature construction process, and more complex algorithms need to be developed to describe the dance movements.

Secondly, the feature set in the proposed DanceTrend framework may be further improved by additional modalities of features. We can further refer to multi-modal analysis methods and add new feature types, such as the audio features, including rhythm, pitch, music type, etc. There are also text features in the title that can be analyzed in combination with the current large language models (LLMs). The LLM-embedded text features may deliver high-level features like emotion status. The cover image characteristics of the published video serve as another feature modality.

Author Contributions: Methodology, S.D., X.H. and F.Z.; Software, S.D., X.H., Y.L., W.Z. and D.F.; Validation, K.L.; Formal analysis, S.D., D.F. and Y.F.; Investigation, S.D., X.H., W.Z., Y.F., K.L., L.H. and F.Z.; Data curation, Y.L. and W.Z.; Writing—original draft, S.D.; Writing—review & editing, Y.F., K.L. and F.Z.; Supervision, F.Z.; Project administration, F.Z.; Funding acquisition, F.Z. and L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Senior and Junior Technological Innovation Team (20210509055RQ), Guizhou Provincial Science and Technology Projects (ZK2023-297), the Science and Technology Foundation of Health Commission of Guizhou Province (gzwkj2023-565), the Science and Technology Project of Education Department of Jilin Province (JJKH20220245KJ and JJKH20220226SK), the National Natural Science Foundation of China (U19A2061), the Jilin Provincial Key Laboratory of Big Data Intelligent Computing (20180622002JC), and the Fundamental Research Funds for the Central Universities, JLU.

Data Availability Statement: The data used and produced in this paper can be divided into three parts. About AIST++, it is openly available at 10.1109/ICCV48922.2021.01315, reference number 40. About Bilibili Dance Data (BDV), which is obtained from the third party Bilibili, is available on request from the corresponding author. The data are not publicly available due to the video contents involves the personal privacy of the publishers, so we cannot provide the original video. We can only provide the id number of the videos we collected.

Acknowledgments: We wish to express our profound gratitude to the three anonymous reviewers for their incisive and constructive comments. Their expert feedback has been instrumental in enhancing the clarity, visualization, and persuasive power of our work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zughear, N.W.; AL-Jabari, M.O. Effects of Using Interactive Web 2.0 Technologies and Mobile Applications on Enhancing Online Shopping Experience for Palestinian Consumers. *Hebron Univ. Res. J. Humanit.* **2023**, *18*, 311–337.
2. Advani, M.; Gokhale, N. Influence of brand-related user generated content (UGC) and brand engagement on Instagram. *AIP Conf. Proc.* **2023**, *2523*, 020105. [\[CrossRef\]](#)
3. Liu, A.-A.; Wang, X.; Xu, N.; Guo, J.; Jin, G.; Zhang, Q.; Tang, Y.; Zhang, S. A review of feature fusion-based media popularity prediction methods. *Vis. Inform.* **2022**, *6*, 78–89. [\[CrossRef\]](#)
4. Hardy, W.; Paliński, M.; Rozynek, S.; Gaenssle, S. Promoting music through user-generated content—TikTok effect on music streaming. In Proceedings of the International 98th Annual Conference, San Diego, CA, USA, 2–6 July 2023.
5. Yuhan, L. Analysis of Body and Emotion in Dance Performance. In Proceedings of the 2021 Conference on Art and Design: Inheritance and Innovation (ADII 2021), Zhengzhou, China, 15 February 2022; pp. 46–50.
6. Oh, C. *K-Pop Dance: Fandoming Yourself on Social Media*; Taylor & Francis: Abingdon, UK, 2022.
7. Panagiotakis, C.; Holzapfel, A.; Michel, D.; Argyros, A.A. Beat Synchronous Dance Animation Based on Visual Analysis of Human Motion and Audio Analysis of Music Tempo. In Proceedings of the Advances in Visual Computing, Rethymnon, Greece, 29–31 July 2013; pp. 118–127.
8. Tsur, O.; Rappoport, A. What’s in a hashtag? Content based prediction of the spread of ideas in microblogging communities. In Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, Seattle, WA, USA, 8–12 February 2012; pp. 643–652.
9. Cui, H.; Kertész, J. Competition for popularity and identification of interventions on a Chinese microblogging site. *arXiv* **2022**, arXiv:2208.10176.
10. Bakshy, E.; Hofman, J.M.; Mason, W.A.; Watts, D.J. Everyone’s an influencer: Quantifying influence on twitter. In Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, Hong Kong, China, 9–12 February 2011; pp. 65–74.
11. Qing, D.; Yefeng, M.; Yi, L.; Hui, Z. Prediction of retweet counts by a back propagation neural network. *J. Tsinghua Univ. Sci. Technol.* **2015**, *55*, 1342–1347.
12. Han, Z.; Tang, Z.; He, B. Improved Bass model for predicting the popularity of product information posted on microblogs. *Technol. Forecast. Soc. Chang.* **2022**, *176*, 121458. [\[CrossRef\]](#)
13. Li, C.; Liu, J.; Ouyang, S. Analysis and prediction of content popularity for online video service: A Youku case study. *China Commun.* **2016**, *13*, 216–233. [\[CrossRef\]](#)
14. Yang, J.; Counts, S. Predicting the speed, scale, and range of information diffusion in twitter. In Proceedings of the International AAAI Conference on Web and Social Media, Washington, DC, USA, 23–26 May 2010; pp. 355–358.
15. Tan, C.; Lee, L.; Pang, B. The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter. *arXiv* **2014**, arXiv:1405.1438.
16. Wang, Z.; Huang, W.-J.; Liu-Lastres, B. Impact of user-generated travel posts on travel decisions: A comparative study on Weibo and Xiaohongshu. *Ann. Tour. Res. Empir. Insights* **2022**, *3*, 100064. [\[CrossRef\]](#)
17. Van Canneyt, S.; Leroux, P.; Dhoedt, B.; Demeester, T. Modeling and predicting the popularity of online news based on temporal and content-related features. *Multimed. Tools Appl.* **2018**, *77*, 1409–1436. [\[CrossRef\]](#)
18. Faridee, A.Z.M.; Ramamurthy, S.R.; Hossain, H.M.S.; Roy, N. HappyFeet: Recognizing and Assessing Dance on the Floor. In Proceedings of the 19th International Workshop on Mobile Computing Systems & Applications, Tempe, AZ, USA, 12–13 February 2018; pp. 49–54.
19. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
20. Fang, H.-S.; Xie, S.; Tai, Y.-W.; Lu, C. Rmpe: Regional multi-person pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2334–2343.
21. Fang, H.S.; Li, J.; Tang, H.; Xu, C.; Zhu, H.; Xiu, Y.; Li, Y.L.; Lu, C. AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 7157–7173. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Yu, B.; Yin, H.; Zhu, Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv* **2017**, arXiv:1709.04875.
23. Markovitz, A.; Sharir, G.; Friedman, I.; Zelnik-Manor, L.; Avidan, S. Graph embedded pose clustering for anomaly detection. 2020 IEEE. In Proceedings of the CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10536–10544.
24. Labuguen, R.T.; Negrete, S.B.; Kogami, T.; Ingco, W.E.M.; Shibata, T. Performance Evaluation of Markerless 3D Skeleton Pose Estimates with Pop Dance Motion Sequence. In Proceedings of the 2020 Joint 9th International Conference on Informatics, Electronics & Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Kitakyushu, Japan, 26–29 August 2020; pp. 1–7.
25. Liu, S.; Fan, Y.; Duan, M.; Wang, Y.; Su, G.; Ren, Y.; Huang, L.; Zhou, F. AcneGrader: An ensemble pruning of the deep learning base models to grade acne. *Ski. Res. Technol.* **2022**, *28*, 677–688. [\[CrossRef\]](#)
26. Lin, S.; Lin, Y.; Wu, K.; Wang, Y.; Feng, Z.; Duan, M.; Liu, S.; Fan, Y.; Huang, L.; Zhou, F. Construction of Network Biomarkers Using Inter-Feature Correlation Coefficients (FeCO₃) and their Application in Detecting High-Order Breast Cancer Biomarkers. *Curr. Bioinform.* **2022**, *17*, 310–326. [\[CrossRef\]](#)

27. Xu, K.; Lin, Z.; Zhao, J.; Shi, P.; Deng, W.; Wang, H. Multimodal deep learning for social media popularity prediction with attention mechanism. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 4580–4584.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
29. Hsu, C.-C.; Kang, L.-W.; Lee, C.-Y.; Lee, J.-Y.; Zhang, Z.-X.; Wu, S.-M. Popularity prediction of social media based on multi-modal feature mining. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 2687–2691.
30. Dunteman, G.H. *Principal Components Analysis*; Sage: Los Angeles, CA, USA, 1989; Volume 69.
31. Abdi, H.; Williams, L.J.; Valentin, D. Multiple factor analysis: Principal component analysis for multitable and multiblock data sets. *Wiley Interdiscip. Rev. Comput. Stat.* **2013**, *5*, 149–179. [[CrossRef](#)]
32. Castro, F.M.; Marin-Jimenez, M.J.; Guil, N.; Pérez de la Blanca, N. Multimodal feature fusion for CNN-based gait recognition: An empirical comparison. *Neural Comput. Appl.* **2020**, *32*, 14173–14193. [[CrossRef](#)]
33. Khosla, A.; Das Sarma, A.; Hamid, R. What makes an image popular? In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Republic of Korea, 7–11 April 2014; pp. 867–876.
34. Gelli, F.; Uricchio, T.; Bertini, M.; Del Bimbo, A.; Chang, S.-F. Image popularity prediction in social media using sentiment and context features. In Proceedings of the 23rd ACM international conference on Multimedia, Shanghai, China, 23–26 June 2015; pp. 907–910.
35. Totti, L.C.; Costa, F.A.; Avila, S.; Valle, E.; Meira, W., Jr.; Almeida, V. The impact of visual attributes on online image diffusion. In Proceedings of the 2014 ACM Conference on Web Science, Bloomington, IN, USA, 23–26 June 2014; pp. 42–51.
36. Chen, Y.-L.; Chang, C.-L. Early prediction of the future popularity of uploaded videos. *Expert Syst. Appl.* **2019**, *133*, 59–74. [[CrossRef](#)]
37. Jeon, H.; Seo, W.; Park, E.; Choi, S. Hybrid machine learning approach for popularity prediction of newly released contents of online video streaming services. *Technol. Forecast. Soc. Chang.* **2020**, *161*, 120303. [[CrossRef](#)]
38. Nisa, M.U.; Mahmood, D.; Ahmed, G.; Khan, S.; Mohammed, M.A.; Damaševičius, R. Optimizing prediction of YouTube video popularity using XGBoost. *Electronics* **2021**, *10*, 2962. [[CrossRef](#)]
39. Sarkar, S.; Basu, S.; Paul, A.; Mukherjee, D.P. ViViD: View Prediction of Online Video Through Deep Neural Network-Based Analysis of Subjective Video Attributes. *IEEE Trans. Broadcast.* **2023**, *69*, 191–200. [[CrossRef](#)]
40. Li, R.; Yang, S.; Ross, D.A.; Kanazawa, A. Ai choreographer: Music conditioned 3d dance generation with aist++. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13401–13412.
41. Wysoczanska, M.; Trzcinski, T. Multimodal Dance Recognition. In Proceedings of the VISIGRAPP (5: VISAPP), Valletta, Malta, 27–29 February 2020; pp. 558–565.
42. Moltisanti, D.; Wu, J.; Dai, B.; Loy, C.C. BRACE: The Breakdancing Competition Dataset for Dance Motion Synthesis. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 329–344.
43. Cheng, T.; Zhang, C.; Chen, G.; Xiao, S.; Zhang, Z.; Jin, X. A Hierarchical Attention-based Contrastive Learning Method for Micro Video Popularity Prediction. In Proceedings of the PACIS 2023, Nanchang, China, 8–12 July 2023; Volume 37.
44. Kuo, T.Y.; Wei, Y.J.; You, B.Y. Chroma Component Generation of Gray Images Using Multi-Scale Convolutional Neural Network. In Proceedings of the 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, 7–10 December 2020; pp. 1243–1246.
45. Wang, Q.; Duan, M.; Fan, Y.; Liu, S.; Ren, Y.; Huang, L.; Zhou, F. Transforming OMIC features for classification using siamese convolutional networks. *J. Bioinform. Comput. Biol.* **2022**, *20*, 2250013. [[CrossRef](#)]
46. Reshef, D.N.; Reshef, Y.A.; Finucane, H.K.; Grossman, S.R.; McVean, G.; Turnbaugh, P.J.; Lander, E.S.; Mitzenmacher, M.; Sabeti, P.C. Detecting Novel Associations in Large Data Sets. *Science* **2011**, *334*, 1518–1524. [[CrossRef](#)]
47. Wang, H.; Li, G.; Wang, Z. Fast SVM classifier for large-scale classification problems. *Inf. Sci.* **2023**, *642*, 119136. [[CrossRef](#)]
48. Nahak, S.; Pathak, A.; Saha, G. Fragment-level classification of ECG arrhythmia using wavelet scattering transform. *Expert Syst. Appl.* **2023**, *224*, 120019. [[CrossRef](#)]
49. Bagher-Ebadian, H.; Janic, B.; Liu, C.; Pantelic, M.; Hearshen, D.; Elshaikh, M.; Movsas, B.; Chetty, I.J.; Wen, N. Detection of dominant intra-prostatic lesions in patients with prostate cancer using an artificial neural network and MR multi-modal radiomics analysis. *Front. Oncol.* **2019**, *9*, 1313. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.