Multimedia Tools and Applications (2023) 82:6965–6991
https://doi.org/10.1007/s11042-022-13583-8

Content-based video recommendation system (CBVRS): a novel approach to predict videos using multilayer feed forward neural network and Monte Carlo sampling method

Baburao Markapudi[1] · Kavitha Chaduvula[2] · D.N.V.S.L.S. Indira[2] · Meduri V. N. S. S. R. K. Sai Somayajulu[3]

Received: 22 January 2021 / Revised: 14 March 2022 / Accepted: 18 July 2022 /
Published online: 11 August 2022

Abstract
Video recommendation has become a crucial role in mitigating the semantic gap in recommending the video based on visual features. This article proposed the exploitation of low-level visual features extracted from videos, and the input data to generate relevant recommendations. Initially, the video is pre-processed with Motion Adaptive Gaussian Denoising Filtering, which eliminates noise from video frames and achieves improved efficiency with high quality and video resolution, which requires less computation. After pre-processing, the paper proposed a content-based extraction approach to retrieve the temporal and spatial characteristics. The temporal characteristics represent the dynamic viewpoints of video, including average shot time and object movement, while the spatial characteristics illustrate a static effect, such as colour and lighting key. Subsequently, this utilizes a series of representative

✉ Baburao Markapudi
baburaompd@gmail.com

Kavitha Chaduvula
kavithachaduvula12@gmail.com

D.N.V.S.L.S. Indira
indiragamini@gmail.com

Meduri V. N. S. S. R. K. Sai Somayajulu
m.somayajulu12@gmail.com

[1] Department of Computer Science and Engineering, Gudlavalleru Engineering College, Gudlavalleru, Krishna District, Andhra Pradesh 521356, India

[2] Department of Information Technology, Gudlavalleru Engineering College, Gudlavalleru, Krishna District, Andhra Pradesh 521356, India

[3] Department of Computer Science and Engineering, Krishna University College of Engineering and Technology, Krishna University, Rudravaram, Machilipatnam, Krishna District, Andhra Pradesh 521004, India

visual features to make the video content more accurate. Finally, the work incorporates a deep neural network to predict the video according to the input and the features extracted. The supervised learning algorithm Multilayer feed-forward is therefore proposed, which generated a series of outputs from a given input set (input data and the extracted features). The majority of deep learning solutions deliver deterministic outcomes and do not measure or monitor prediction variance, which can contribute to a loss of faith in automatic evaluation. Subsequently, Monte Carlo's uncertainty techniques are used to estimate the exact video in accordance with the recommendation. The proposed method is implemented using MATLAB r2020a software with less computation time of 0.999 s and the performance of the proposed method is compared with the different existing methods like MMM, LP-LGSN, and $CDPR_{ec}$. Consequently, the proposed method produces higher performance in terms of precision, recall, F-measures, and nDCG and it produces higher accuracy of 0.94%, respectively.

Keywords Video recommendation · Motion adaptive Gaussian Denoising filtering · Content-based extraction technique · Visual features · Multilayer feedforward neural network · Monte Carlo sampling

1 Introduction

A daunting volume of video content from blog sites and video sharing are faced by online users [27]. As a result, the user is watched or watching a movie, the user's most attracted movie is concluded by the Video recommendation system and this system recommends these movies to the user [22]. To pick out the favourite video, this system saves users from browsing many movies and it brings user stickiness and video websites are more network traffic. Recommendation services are already provided to Netflix, YouTube, Google Video, and Yahoo!, these are the existing video-oriented sites, most of which likely recommend videos [5]. For a better recommendation, clicking and leveraging video content remains a difficult challenge for the consumer. The recommendation systems describe the selection of objects that are more enticing and important to the consumer and the potential to sort vast recommended systems [20]. Hybrid filtering, Collaborative Filtering, and Content-based filtering are important parts of the video recommendation algorithm. The prominent part of the video recommendation system is the video recommendation algorithm. To compute recommendations and to build the user model, Content-based Filtering models use the preference indications of content information available about the items and a single target user [25]. On existing approaches, video content is automatically analyzed and a set of representative features are extracted using a technique is encompassed by the Content-based recommender system [26].

The content-based extraction technique extracts the visual features of the videos. The video items are suggested by the popular content-based methods and these content-based methods have similar content characteristics to the user-liked items in the past. To find similarities in articles, terms or words are considered by the news recommendation as an example of Content-based methods [4]. The availability of data on the related content features of objects for content-based filtering is a prerequisite [8]. Unstructured or structured meta-information these features are related to the video items, in most existing systems. Directors consider movie genre, textual reviews (unstructured information), structured information, tags, plot, and cast are examples of many recommendation systems in the movie domain [19]. To improve their accuracy, the explicit contents are exploited by the other content-based techniques. Low-Level features (LL) and High-Level features (HL) are the two types of item features typically exploit by the recommender systems in the multimedia domain

[6]. News articles, reviews, social tags, and item descriptions are the less structured data or ontologies, lexicons and databases are the structured sources of meta-information these are used to obtain media content's High-Level features express properties. Low-Level features are directly extracted from media files themselves [1].

A prediction method is used to predict the relevant videos based on their visual features and user input. The stylistic aspect of the video is typically represented by the Low-Level features and the uses of Low-Level features were investigated by a limited amount of works in the video recommendation domain [7]. However, existing methods only take into account scenarios where, in addition to another category of knowledge, exploit the low-level features to enhance the recommendation quality. In order to enhance the click-through rate, a video recommender system is proposed [3] and it is called Video Reach which uses a combination of Low-Level and High-Level video features of different aural, visual, and nature - textual. Various visual content and data sources are used to generate multiple ranking lists; it is integrated by a multi-task learning algorithm [11]. High-level information is missing during the video files only available, only Low-Level visual features are used in none of the work when the extreme new item problem occurs [16].

This research focused on predicting the video based on their extracted features and the user input for a recommendation but the existing techniques only focused on feature extraction for video recommendation. Based on the user request and visual feature extraction techniques, this Multilayer feedforward neural network accurately predicts the video. The features of the video are extracted by using the Content-Based Extraction technique and the relevant video is predicted by using the deep neural network. Then, the uncertainty method of the Monte Carlo sampling method helps to classify the hard samples of output. The deep neural network simulator program is developed using MATLAB software. Below section 2 depicts the literature survey, the problem definition and motivation are portrayed in section 3. Section 4 explained the research methodology, experimentation, and result discussion are explained in section 5, and section 6 explains the conclusion.

2 Literature survey

This survey is based on the video recommendation system for predicting accurate videos based on the video's visual features. A Multilayer feed forward Neural Network-based uncertainty estimation technique is proposed to accurately predict the video.

Pu, Shi, et al. [15] proposed a multimodal topic learning algorithm to exploit three modalities (i.e., cover images, tags, and titles) for generating offline video topics. The proposed algorithm generates semantic topic features, which help with preference recommendation generation and scope determination. The online computational cost is effectively reduced instead of using visual content features, semantic topic features are used. Consequently, the presented algorithm has been implemented in the Kuaibao information streaming platform. The proposed algorithm performs favourably as shown by the offline and online evaluation results.

Wang, Xichen, et al. [23] proposed a content-based recommendation algorithm Category-aided Multi-channel Bayesian Personalized Ranking (CMBPR). The difference among both various video and various user interaction categories is considered, the rich preference information of the user is integrated by the short video recommendation. The video recommendation algorithm's effectiveness for CMBPR was demonstrated by the experimental results, compared to the traditional video recommendation algorithms, it achieves a

significantly higher recommendation accuracy, and the "Long Tail" effect influence is solved by the proposed algorithm.

Mehta, et al. [14] presented a transform-based approach for denoising images corrupted by Gaussian noise using wavelets. Five wavelets were chosen after an extensive study of wavelets on a wide variety of benchmark images. Obtain a denoised output using the modified Bayes thresholding for each wavelet. To obtain the final denoised image, the median of every pixel is found from these outputs. In terms of both quantitative and qualitative efficacy measures, the output obtained from this approach outperforms other methods.

Tong Wua et al. [24] suggested a customized UGC video utility recommendation framework under an interest graph of the entire network, creating and incorporating a user interest chart, user interest progression, and reviews. Analysing and recognizing profoundly educated video content. The core hypotheses and development strategy for the video suggestion method were thoroughly researched based on user-interest graphs. The aim is to create a customized recommendation system for a user-interest graph and UGC video resources and to provide solutions for customized network video resource recommendations.

Hilman Fauzi Rija et al. [9] describe the Indonesian AVSR system (INAVSR) syllable development by the audio and visual feature's fusion. Both this PCA (principal component analysis) and DCT (discrete cosine transform) are used to extract the visual feature and Hidden Markov Toolkit (HTK) is used to develop the system. The acoustic model based on phonemes and letters/graphemes/characters are commonly not robust to noise by creating an optimum combination of visual features and audio and feature extraction technique is Gaussian Mixture Model (GMM) and MFCC (Mel Frequency Cepstral Coefficient) and it has high complexity.

Tippaya, et al. [21] analyzed the visual representation behaviour in terms of the discontinuity signal by presenting a multi-modal visual features-based SBD framework. The use of discontinuity signals average cumulative moving but without the threshold is performed by adopting a selection of candidate segment, video frames for non-boundary are ignored and shot boundaries location is identified. The logo occurrence and fade in/out are included a gradual transition and a cut transition into this candidate segment is distinguished, the transition detection is structurally performed.

The aesthetic model for characterizing films with a multidisciplinary approach has been presented by lvarez et al. [2]. Combining movie theory, visual low-level video descriptors, and classification techniques using machinery and profound learning. Four various tests for different applications have been developed that prove the utility of the model. The applications are production year prediction, aesthetic clustering style, film popularity influence, and genre detection. The results are compared to high-level data to assess the accuracy of the model for classifying films without previous knowledge.

In Heterogeneous Knowledge Networks (HINs, for its excellent characterization of heterogeneous and dynamic background information, Lei sang et al. [17] examined the issue of video suggestion. Accordingly, a CDPRec network has been suggested to accurately video input and collect global contextual references for HIN images. In a graphical HIN and the surrounding video nodes, CDPRec may propagate video context along with a series of links and explore several dependencies. The composition of the multimodal content function and the global knowledge on dependence structures is expressed by an attentiveness network in each video. The learned video integration and the sequential recommendation are optimized jointly for the final rating forecast.

The entire R&D prediction system was introduced by Suiyi Ling et al. [10] by identifying discriminating features. A first Bjontegaard Delta (BD)-Rate, a BD-Quality-driven algorithm is

proposed to categorize UGC to improve understanding of the UGC behaviours. By using a hierarchical selection framework for the R-D related category as a ground truth label, further identifying features characterizing UGC R-D behaviour. Finally, selected characteristics are used to forecast the R-D subtest UGC category.

3 Problem definition and motivation

The user to choose the right video is provided by the video recommendation system, it is an efficient way for achieving greater user loyalty and stickiness of consumers. Therefore, scholars and video websites pay much attention to it. The user watching or watched a movie is analyzed by the video recommendation system, and this recommends the movie to the user which movies may attract the user most. Subsequently, this system helps the user from browsing many movies select favourite ones then it brings user stickiness and more network traffic in video websites. Additionally, the video is automatically recommended from the like videos. Subsequently, it predicts some unwanted videos to the user or it predicts not related videos instead of predicting user recommend video. It is one of the major problems faced by the user. Therefore, to get the user recommended video, the prediction is occur based on the user request input and the visual features and it gets the most relevant video to the user.

In online services, the video recommendation system is the most popular one. The present invention detects Gaussian noise in videos and it also detects the artifacts in a Video frame. Based on this abundant noise, the Low-resolution videos are usually corrupted. The noise in the videos is removed by several filtering techniques. The noise extent practicable is diminished based on the temporal filter and it should be used in the frames motion area by the video noise reduction algorithms. Visual feature extraction is computer vision's key technology for intelligent video processing. For reliable object detection, the stable visual feature points are provided by the SIFT (scale-invariant feature transform) and it is the most commonly adopted approach among extraction techniques. Additionally, among several sensor nodes, the processing task is distributed in the network by real-time processing of the visual information that is allowed by some techniques. One of the challenging tasks is optimizing the distribution of the processing tasks among the network nodes. Therefore, a content-based extraction technique is proposed in this research, the feature of the video is extracted automatically and the time is reduced. In recent decades the prevailing techniques utilized various techniques for accurate prediction, Speech recognition, computer vision, and natural language processing are the application fields for this application, and deep learning has witnessed great success.

4 Research methodology

Recommendation system (RS) has been paid attention based on the rapid development of mobile internet. YouTube, yahoo, Netflix, and Amazon are well-known e-commerce platforms that are gradually equipped with a Recommendation System. With the emergence of online social networks (OSNs), video recommendation has become increasingly important in mitigating the semantic gap. Motion Adaptive Gaussian denoising Filtering is a noise removal technique, the noises are removed from the video and give a high quality and clarity video frame and this is the pre-processing step with less computation requirement. Figure 1 demonstrates the architecture of the proposed methodology.

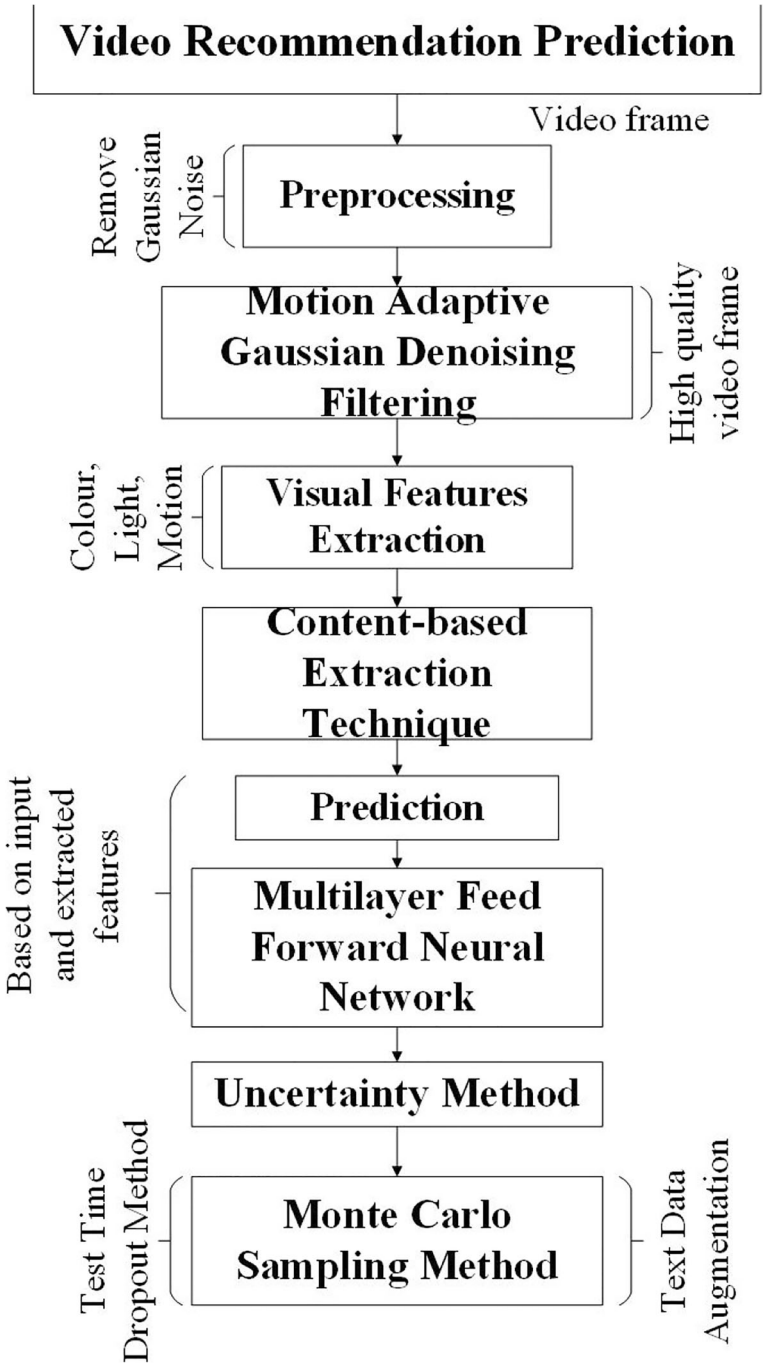


Fig. 1 Architecture of Proposed Methodology

The content Based Extraction technique is introduced to extract the visual features of the video after completing the pre-processing technique. Colour variation, object motion, and

lighting key are the extracted features of the video. After the features are extracted, the Multilayer Feedforward Neural Network prediction the video based on these visual features and user input. Further, the uncertainty problem is overcome by the Monte Carlo sampling method, which adopts two techniques like test time dropout and test data augmentation method to generate the accurate video. The overall process is explained in the following sections.

4.1 Motion adaptive Gaussian Denoising filtering (MPGDF)

Gaussian noise or unwanted noises from the videos are removed by the Motion Adaptive Gaussian denoising Filtering and it gives high-quality videos it is a pre-processing step. According to an optimization algorithm, variable parameters are adjusted and control the transfer function based on the adaptive filter it is a system with a linear filter. Because of the optimization algorithm's complexity, all the adaptive filters are digital filters. According to the shape of the Gaussian function, it is a linear smoothing filter that chooses the weights. To remove the noises a Gaussian smoothing filter is an effective low-pass filter that is subject to the normal distribution whether in the frequency domain or the spatial domain. In image processing, it has a broad prospect of application. Equation of zero mean one-dimensional Gaussian function is given in Eq. 1:

$$f(a) = e^{\frac{a^2}{2\sigma^2}} \tag{1}$$

Where, σ is the width of the Gaussian function and the Gaussian function is a single value function. To replace the current pixel, the pixel neighbourhood's weighted mean is used by the Gaussian filter, and the centre point becomes far away. To reduce noise or to blur the image a Gaussian filter is usually used and this filter is linear. The second derivative of the Gaussian filter is determined to obtain the coefficients of the Gaussian filter. The below equation is an ID Gaussian equation,

$$f(a) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(\frac{-a^2}{2\sigma^2}\right)} \tag{2}$$

Where, σ- Standard deviation of Gaussian filter and a - Gaussian index and the Gaussian filter's first derivative equation is given as follows.

$$f^{'}(a) = \frac{-1}{\sqrt{2\pi}} \frac{a}{\sigma^3} e^{\left(\frac{-a^2}{2\sigma^2}\right)} \tag{3}$$

The Gaussian filter's first derivative equation is shown in the following eq. (4),

$$f^{''}(a) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma^2} \left(\frac{a^2}{\sigma^2} - 1\right) e^{\left(\frac{-a^2}{2\sigma^2}\right)} \tag{4}$$

The rise time and fall time are minimized to a step function input that the Gaussian filters have the properties of having no overshoot.

4.1.1 Gaussian Denoising filter

For Gaussian denoising, non-local mean filters are utilized. An algorithm called iterative non-local means (NLM) is used to improve the corrupted pixel's value approximation after

replacing and identifying the corrupted pixels. According to its intensity and that of its neighbours, change each pixel's intensity and the spatial domain defined a spatial filter. The input non-linear function dictates the output of non-linear filters. The total variation is assumed by the total variation methods and it has a high signal gradient integral is consequently assumed. Obtain a higher similarity denoised video by decreasing total variation. In the denoising algorithm, the total variation is minimized to reconstruct the original video from a noisy one by using the Motion Adaptive Gaussian Denoising Filtering.

$$\min \frac{\left\|X-i^2\right\|}{2\lambda} + Y(X) \tag{5}$$

Where, the clean image is denoted as X, the observed image is represented by i, the larger multiplier as λ, and $Y(X)$ is the total variation. Patch processing and optical flow combine motion estimation by using the proposed denoising algorithm. Before the quantization stage, the high-frequency content is reduced at low bitrates to help the encoding stage by the adaptive low-pass filters in the presence of noise the video sequences are not altered, in video professional applications it is a common situation. For high-quality applications, due to the annoying effect of smoothing the adaptive filters suffer from this effect. Based on the number of display parameters, the proposed pre-processing filter has the particularity of depending. The total variance $Y(X)$ is represented in the following equation,

$$Y(X) = \sum_{(x,y)=1}^{N} \left|(\nabla X)_{x,y}\right| \tag{6}$$

Mainly for high-frequency sub-bands, the remaining noise is removed and smooth the image by applying an improved total variation. The high-quality videos or clarity videos are given by the proposed Motion Adaptive Gaussian Denoising Filtering (MPGDF). The quality video is pre-processed by a motion model that quantifies motion coherency. In the pre-processing step, the Gaussian noise is reduced and a high-quality video is given by this proposed MPGDF.

4.2 Content-based extraction technique

The Content-Based Extraction technique (CBET) extracts the spatial and temporal features these are the features videos. After the pre-processing step, extract the visual features of the video, that is shot duration, object motion, lighting key, and colour variance are the temporal and spatial features of the videos. Keyword-based models represent the content, in which item features representation is Vector Space Model (VSM) are created by the recommender, wherein a multidimensional space vector represents an item. The items are described by the dimensions representing the features. A relevance score is measured by the system and the user's degree of interest toward any of these items is represented.

4.2.1 Recommendation system

The features with High-Level (HL) and Low-Level (LL) are the two types of features in the recommendation system. News articles, social tags, item descriptions, and reviews are the less structured data or lexicons, databases and ontologies are the structured sources of meta-information from these obtain a media content's High-Level features express properties. The media files directly extract the Low-Level features. Timbre or rhythm are the acoustic

properties it is a Low-Level feature, to discover video tracks that are similar to user liked video that features are exploited. The use of Low-Level features is investigated by a limited number of works, and the pure visual contents extract the features, stylistic aspect of the video is typically represented by this in the video recommendation domain. To improve the quality of recommendations, another information type exploits the Low-Level features that the existing approaches consider only scenarios.

The video's visual features are extracted by the proposed Content-Based Extraction technique (CBET). A classical content-based algorithm is adopted utilizing a Low-Level stylistic visual features recommendations are generated. Given a set of preference scores g_{ui}, a catalog of items is represented as $i \in I$, and a set of users is represented as $u \in U$ is given by collecting an item i from a user u. In the b_i feature vectors, associate each item $i \in I$. Cosine similarity is utilized to compute the similarity score a_{ij} for each couple of i and j items.

$$a_{ij} = \frac{{b_i}^T b_j}{\|b_i\| \|b_j\|} \tag{7}$$

For each item, i the set of its nearest neighbours NN_i is built, $|NN_i| < K$. For an unseen item i is computed, the predicted preference score $\widehat{g}_{ui}$ for each user $u \in U$ is given as follows;

$$\widehat{g}_{ui} = \frac{\sum_{j \in NN_i, g_{uj} > 0} g_{uj} a_{ij}}{\sum_{j \in NN_i, g_{uj} > 0} a_{ij}} \tag{8}$$

The first step to build a video CBET based on low-level features is to search for features that comply with human visual norms of perception. The focus of this work is only on visual features, therefore is a set of features that describe the visual content of a video. A video can be thought of as a contiguous sequence of many frames in generally. Subsequently, several frames are quite similar and associated in consecutive video frames. Considering all of these frames for feature extraction not only does not provide new information to the system, but it is also inefficient in terms of computing. Consequently, a structural analysis of the video is performed before feature extraction, which includes detecting shot borders and extracting a keyframe within each shot. A shot boundary is a frame in which the visual content of the frames around it differs significantly. On the other hand, frames within a shot are so similar, it makes sense to select one sample frame from each image and utilise it for feature extraction.

4.2.2 Visual features

Histogram-based, a motion vector, multiple features-based, temporal, pixel-based, descriptor-based compressed domain feature, spatial feature-based, combined feature-based, and edge-based are the different groups that are categorized by the CBET scheme and this is the use of extraction of visual feature. Each video frame's low-level feature is extracted by the visual information of the video frame represented by one common approach. The image without distinct regions or shape information (spatial relationship) is described by these features. Local features and global features are the two main types of features are extracted features. From the colour histogram that can be used, the global feature difference is extracted for a CBET system. To a movement of the object or small camera, histogram colour is less sensitive due to the property that the colour variation's spatial information does not incorporate. However, due to rapidly changing colour details, in an object motion or large camera differentiate two shots within the same scene is incapability it is more sensitive and these are their disadvantage.

Compare to the global features, local features movement is small and more tolerant of illumination changes, but computational complexity is higher. Based on the properties of images, the above-mentioned features are mainly relied on. Motion vector and dominant colour are the low-level features used in a video, it constructs the middle-level features. To map the raw feature onto the smaller dimensional vector, while preserving the temporal characteristic of video content, propose some dimensionality reduction techniques. From each video, the most distinctive and informative features are extracted.

$$v(z) = \left(\overline{X}_S, \overline{X}_M, \overline{X}_{\left(\sigma_m^2\right)}, \overline{X}_C, \overline{X}_L\right) \tag{9}$$

Where the average of shot length is $\overline{X}_S$, $\overline{X}_M$ is average object motion, $\overline{X}_{\left(\sigma_m^2\right)}$ is the standard deviation across all frames respectively, $\overline{X}_C$ is the average colour variation, and $\overline{X}_L$is the average lightning key. To convey emotions each feature carries meaning and in the hand of an able director, it is utilized.

Shot duration A movie is being created by the pace and a shot is considered a single-camera action it is useful information that is given by the number of video shots. The average shot length $\overline{X}_S$ is given as follows,

$$\overline{X}_S = \frac{f_n}{s_n} \tag{10}$$

Where the number of frames is represented as f_n and s_n is the number of shots in the video. Different frame rates are present in the videos, the video frame rate is normalized $\overline{X}_S$.

Object motion A single key point represents image features whose motion is highly correlated are likely to belong to the same object. Therefore, a separation loss is added that promotes keypoint trajectories to be decorrelated in time. Independent object motion is represented by one particular pattern type it is a motion boundary-based heuristics. Filmed the motion on part of the object. The former characteristic of a movie is captured by the average shot length and it is desired that the latter characteristic is also captured for the motion feature. At the frame$r$, $\overline{m}_r$ represent the average motion of pixels and $(\sigma^2{}_m)$ r is the pixel motion's standard deviation:

$$\mu_{\overline{m}} = \frac{\sum_{r=1}^{f_n} \overline{m}_r}{f_n} \tag{11}$$

$$\overline{X}\left(\sigma^2{}_m\right) = \frac{\sum_{r=1}^{f_n} \left(\sigma^2{}_m\right) r}{f_n} \tag{12}$$

Where the average motion means is $\mu_{\overline{m}}$ and the motion standard deviation is $\overline{X}_{\left(\sigma_m^2\right)}$ aggregated over entire frames f_n. Filmed the velocities of images, in the image sequence a robust estimate of the motion is measured based on optical flow, a motion feature descriptor used. Across all video frames, motion features are calculated because features of motion depend on the image sequence.

Colour variance To the lighting colour, expressive quality is closely related; given the situation derive the feeling is magnified or shares the same ability to set. The challenge with colours is isolating their contribution to a scene's overall "mood" from other aesthetic variables operating in the same sense. Toward a specific emotional objective, predispose the context as a whole that their effectiveness is higher.

Enough scientific data not currently supported the feeling and colours correlation that may be evoked investigate colours that nonetheless have an expressive impact. Based on the genre, the colour variance has a strong correlation. For videos, a large variety of bright colours used for directors tend for instance. The covariance matrix is computed and in Luv colour space each key-frame is represented.

$$\tau = \begin{bmatrix} \Sigma_L^2 & \Sigma_{LU}^2 & \Sigma_{LV}^2 \\ \Sigma_{LU}^2 & \Sigma_U^2 & \Sigma_{UV}^2 \\ \Sigma_{LV}^2 & \Sigma_{UV}^2 & \Sigma_V^2 \end{bmatrix} \tag{13}$$

In each keyframe, for the representative of the colour variance, the generalized variance can be used, and the equation for the keyframe is given as follows,

$$\sigma_p = \det(\tau) \tag{14}$$

In the middle shot, the keyframe is a representative frame, and then calculate the average colour variance is given as Eq. (15),

$$\overline{X}_C = \frac{\sum_{p=1}^{s_n} \sigma_p}{f_n} \tag{15}$$

Where, s_n is the number of shots equal to several keyframes. The colour covers in an image, brightness, colour's saturation, and the area size based on this a quantity that depends. The hue plays a role when the quantity of energy is less hue tends to more blues and it has more energy quantity when it tends toward reds.

Lighting key Between the video genres, another distinguishing factor is the lighting key in that way emotion type is controlled when they want to be induced to a viewer; the director uses it as a factor. High gray-scale standard deviation is the dimmest light and high gray-scale mean is a light abundance with less contrast between the brightest and dimmest light, a lighting key is often adopted by the comedy videos. Consequently, this trend is often known as high-key lighting. In both gray-scale standard deviation and gray-scale mean, horror videos often pick gray-scale distributions which are low and it is known by low-key lighting. To capture both of these parameters, compute the standard deviation σ and mean μ of the value component, to HSV colour-space all key-frames are transferred after that it corresponds to the brightness. The lighting of keyframes is measured by using the multiplication of μ and σ it defines the scene lighting key ξ,

$$\xi_p = \mu.\sigma \tag{16}$$

In both the mean and standard deviation of gray-scale values, comedies often contain key-frames that have a well-distributed gray-scale distribution. One can state $\xi > T_a$ for the comedy genre, whereas the lighting key with poorly distributed lighting the situation is reversed $\xi < T_b$ for horror movies, were T_a and T_b are the predefined thresholds. The above

distinguish factor is hard to use for them in a situation where $T_b < \xi < T_a$ other video genres exist. The following eq. (17) is the average lighting calculated over keyframes,

$$\overline{X}_L = \frac{\sum_{p=1}^{s_n} \xi_p}{s_n} \tag{17}$$

By using this implementation, extract the low-level visual features.

In terms of low-level visual features, the frames of the videos can be reflective of their full-length videos, indicating a significant correlation with them. After extracting the features of the video, feed-forward supervised learning is utilized which is depicted in the following subsection.

4.3 Multilayer feed forward supervised learning technique

The content-based feature extraction technique extracts the data; these extracted data are given to the neural network's input layer. Based on the extracted features, the feed-forward network is employed to analyse and predict the videos. One of the most basic artificial neural networks is the feedforward neural network. The data or input presented in this ANN ravels in a single direction. The data enters the ANN via the input layer and exits via the output layer while hidden layers may or may not exist. Consequently, the feedforward neural network rarely has a backpropagation but it only has a front-propagated wave. There can be multiple hidden layers that depend on what kind of data you are dealing with. The number of hidden layers is known as the depth of the neural network. More functions can be learned by the deep neural network. The proposed article utilized two hidden layers to generate the set of outputs with the given set of inputs. The input layer feeds data into the neural network, while the output layer produces predictions based on a sequence of functions, which is depicted in Fig. 2.

Input variables are represented by the input layer's neurons, and the predicted output value is represented by a single neuron in the output layer. The hidden layer consists of two hidden units, it is used for cross-validation of the prediction model and the hidden layer performs computational tasks. For these two hidden layers, one hidden layer is used for training, and the other hidden layer is used for validation. All nodes of the input layer are connected to the hidden layer's node and all nodes in the hidden layer are connected to each node of the output layer.

A nonlinear function is $x : P \in A^Z \rightarrow Q \in A^1$ it is generalized by the network for function modelling purposes with one predicted variable. In Eq. 18, the function for x is succinctly explained.

$$x(P) = v_2 + R_2 \times (x_K (v_1 + R_1 \times P)) \tag{18}$$

The hidden layer's weight matrices are R_1 and R_2, and the hidden layer's bias vector is denoted by the $v_1 = [v_{11}, v_{12}, \ldots, v_{1N}]$ is the output layer; the output layer's bias vector is v_2; an activation function is represented by x_K.

Initialize the weight matrices R_1 and R_2, and the bias vectors v_1 and v_2, and then updated through a training process. Based on their experience training epoch, momentum, and learning rate were set these are the training parameters. Between the corresponding connection weights and the input neuron values, each hidden node's value from the sum of multiplications is

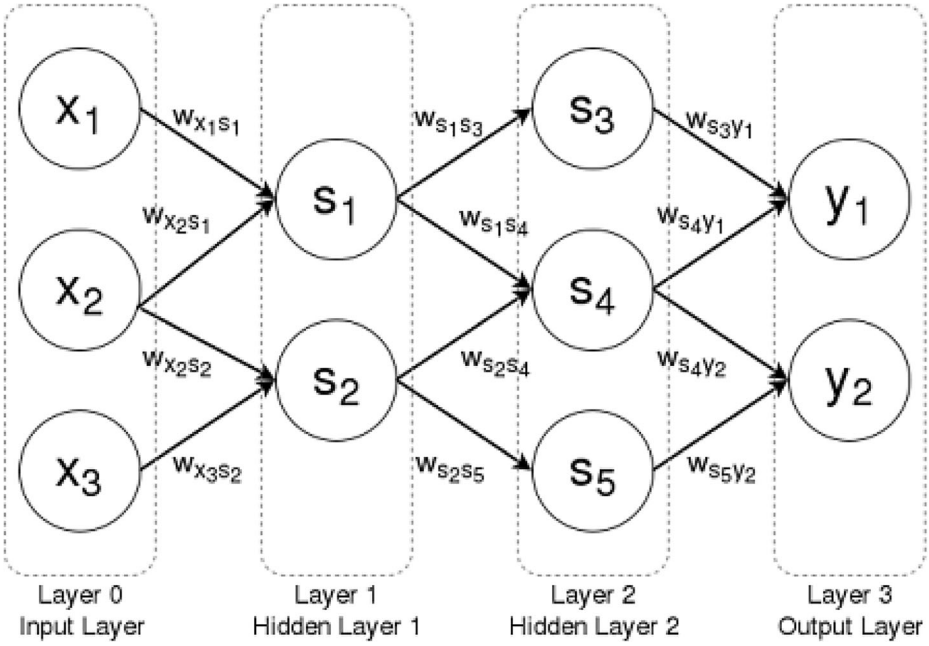


Fig. 2 Feed Forward Neural Network

evaluated in the forward stage. From connection weights and the values of the hidden nodes, in the same way, the output value is propagated forward. Consequently, by comparing the output response values to target values, the difference that can be minimised in the backward stage is estimated by updating the connection weights. Most approaches do not quantify the uncertainty in predictions and are unable to recognise data that is abnormal or significantly different from that used in training, which can lead to a lack of confidence in the automated forecast and its interpretation. Consequently, the work explores the use of Monte-Carlo sampling to solve the uncertainty problem which is explained in the following section.

4.4 Monte-Carlo sampling method

The Monte-Carlo Sampling method was utilized to overcome the uncertainty problem in the deep learning technique. The technique utilized two approaches for the prediction of accurate video such as test time drop out and test data augmentation. Invoke the Test Time Dropout approach by random removal of model uncertainty (with a dropout rate of 0.2).

The drop-out at the test time uses the same likelihood to monitor model uncertainty for a given video frame p and perform many predictions $\{P_t\}_{t\ =\ 1\ldots T}$, each prediction is an A-class softmax scoring vector. Additionally, employs the test data augmentation technique of prediction uncertainty. The videos are forwarded multiple times during the evaluation time via the neural network with random data enhancement setups.

Receive versions by adding data for each image p and transmitting the frames to obtain $\{x_t\}_{t\ =\ 1\ldots T}$ a number of predictions $\{P_t\}_{t\ =\ 1\ldots T}$. Finally, combine these two approaches to estimate both uncertainties: generate additional examples T for each frame p and forward each example by drop-out at test time. The methods are used in-depth below to analyse the uncertainty measures.

Table 1 Types of Videos Used for Training

Video Types	Number of Videos
Animation Videos	12
Cover Songs	18
Gaming Videos	16
Lecture Videos	16
Live Music	16
Lyric Videos	5
Music Videos	13
News Click	24
Spots Video	24
Television Click	6
Vertical Videos	20

Entropy Entropy is an average knowledge level or variance resulting in the potential results of a random variable.

$$E(P_T(p)) = -\sum_{a=1}^{A} P_T(p)[A]\log(P_T(p)[A]) \tag{19}$$

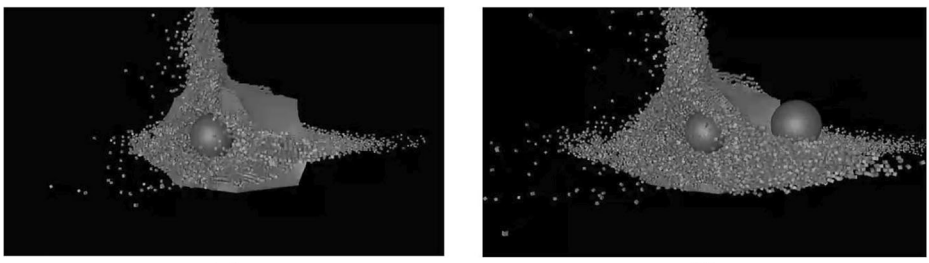
Where, $(P_T(p)[A])$ indicates the A^{th} element of the vector and $P_T(p)$ exposed class (A) average prediction score.

Variance Compute the difference between forecasts and mean variances T in all classes within each class.

$$\sigma^2(P_T(p)) = \frac{1}{T}\sum_{t=1}^{T}(P_t(p) - P_T(x))^2 \tag{20}$$

Bhattacharyya coefficient (BC) The coefficient of Bhattacharyya is a measure of the overlap between two statistical samples or populations. Calculate the generalized BC with a higher predictive average for the two groups.

$$BC(q_{a1}, q_{a2})(x) = \sum_{n=1}^{N}\sqrt{q_{a1}[n] * q_{a2}[n]} \tag{21}$$



(a) **(b)**

Fig. 3 Animation Video Frames

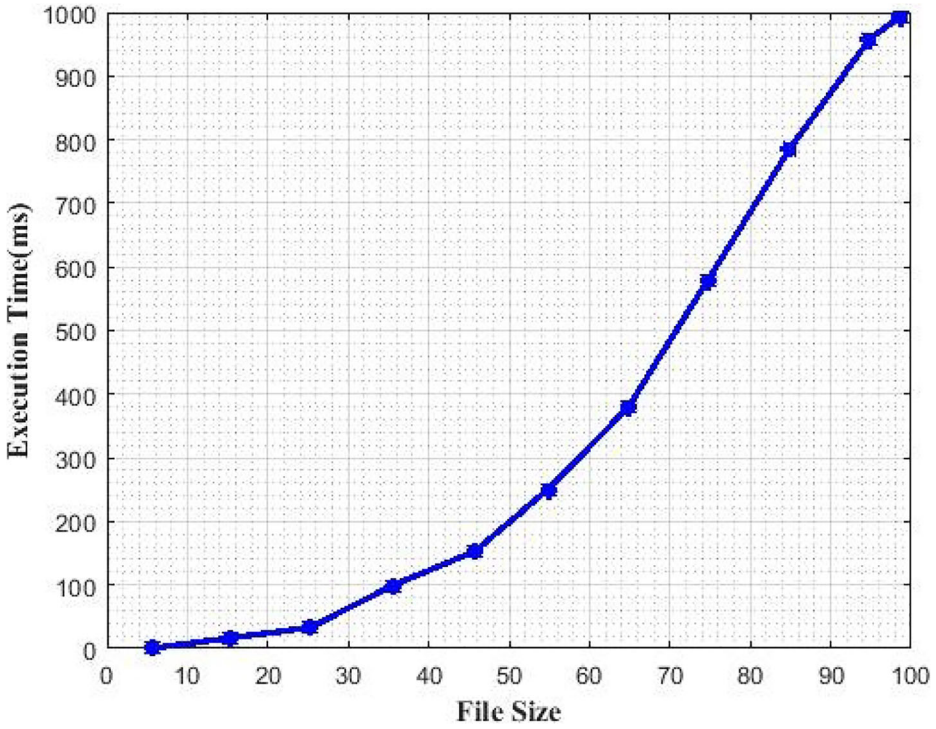


Fig. 4 Execution Time of Proposed Method

Where, q_{a1} *and* q_{a2} indicates the N-bin histograms of the two classes with higher predictive mean $P_T(x)[a1]$ *and* $P_T(x)[a2]$. These are the measures that predict the accurate video in terms of balanced accuracy in uncertainty measurement.

5 Experimentation and result discussion

The data are collected from the standard UGC dataset; data are in the form of videos; therefore 170 videos are collected from the UGC dataset. The performance evaluation has been carried out with the input set of 170 videos for training. Consequently, the 170 videos are in the form of animation videos, cover songs, gaming videos, lecture videos, live music, lyric video, music video, news click, sports video, television click, and vertical videos, which is depicted in Table 1.

Table 2 Simulation System Configuration

MATLAB	Version R2020a
Operation System	Windows 10 Home
Memory Capacity	6GB DDR3
Processor	Intel Core i5 @ 3.5GHz
Simulation Time	10.190 seconds

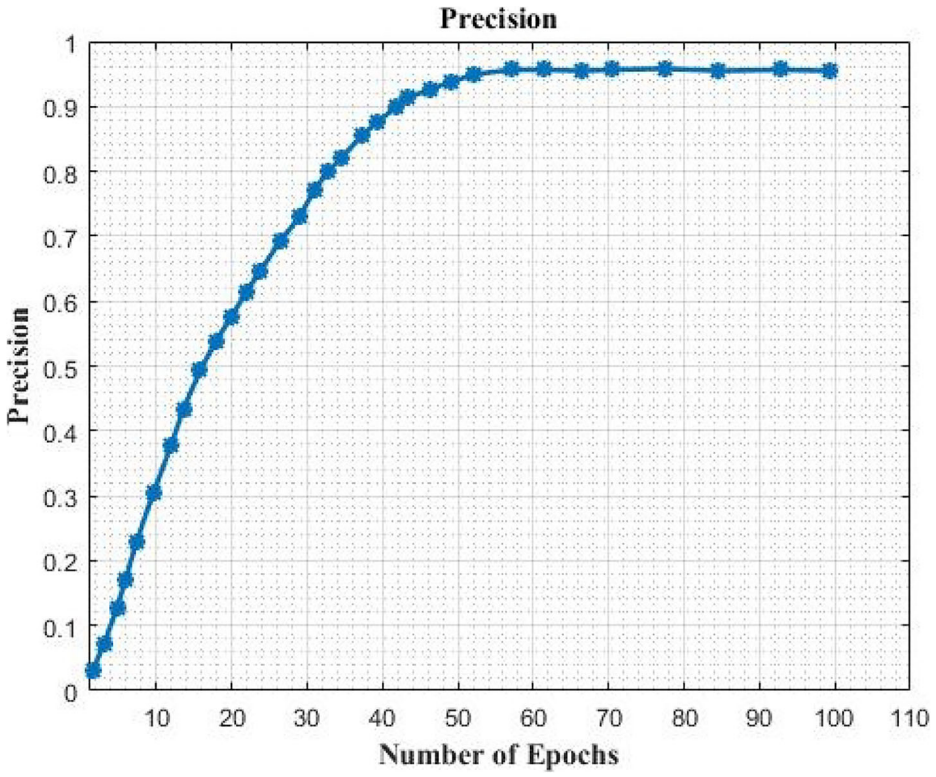


Fig. 5 Precision Graph of Proposed Method

Subsequently, the collected videos consist of 12 animation videos, 18 cover songs, 16 gaming videos, 16 lecture videos, 16 live music, 5 lyric videos, 13 music videos, 24 news clicks, 24 spots videos, 6 television clicks, and 20 vertical videos. These videos are processed by using the proposed techniques.

Above Fig. 3(a) and (b) are frames 1 and 2 of animation videos for prediction based on this neural network. The picture is one of many still images that make up the complete moving image. Analog waveforms were represented to display video frames, with varying voltages representing light intensity in an analog raster scan across the screen.

5.1 Simulation output

Figure 4 describes the execution time for this proposed methodology. The task's execution time refers to the time spent by the system performing a task, including time spent by the neural network performing run-time or system services on its behalf. The proposed technique takes time as 0.999 seconds for execution.

The proposed technique has been implemented using MATLAB software. Table 2 describes the system configuration for simulation, this method has been tested and evaluated by using the MATLAB r2020a. Operation System for this software is Windows 10 Home and its memory capacity is 6GB DDR3. Intel Core i5 @ 3.5GHz is the Matlab processor and the time required for simulation is 10.190 seconds.

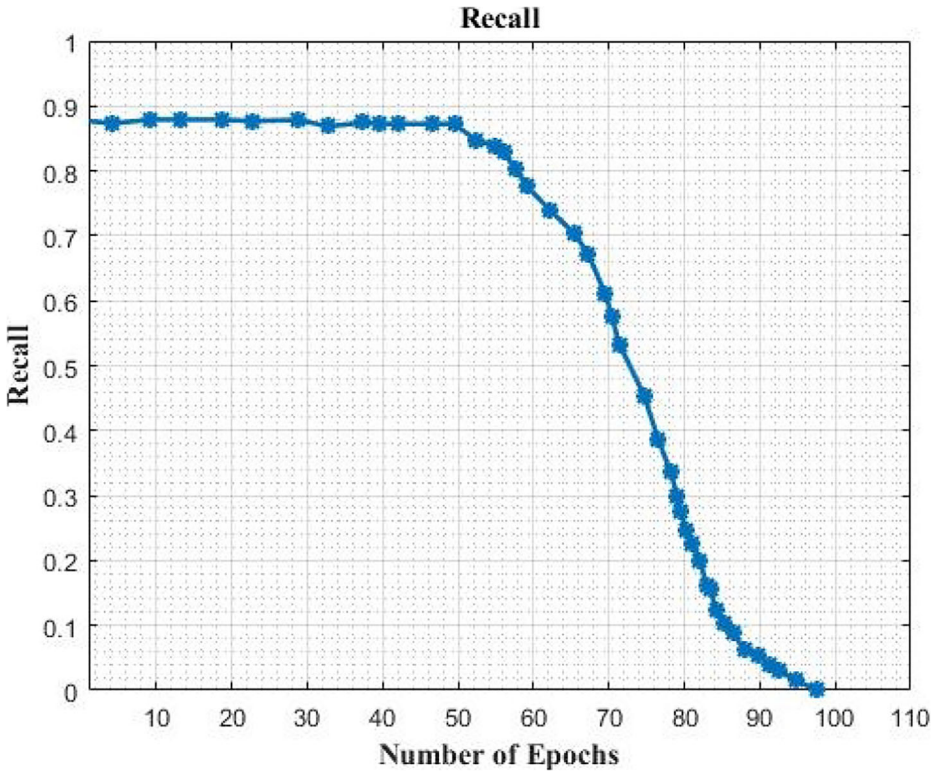


Fig. 6 Performance Graph for Recall

5.2 Performance analysis

In order to evaluate the recommendation results, train, and test the proposed model. In this experimental section, the precision (P), recall (R), F-score (F), and nDCG are calculated, these are the evaluation metrics of this proposed method. The precision determines the number of positive class predictions that belong to the positive class. Precision can be evaluated by using the following formula,

$$\text{Precision}\,(P) = \frac{N_k}{k} \tag{22}$$

The recall measured the number of positive class predictions made out of all positive cases in the dataset. Where N is the number of ground truth videos of each user.

Figure 5 depicted the performance graph of precision, it depicted that the precision increases with increasing the number of epochs. When the number of epochs increases from 10 to 100, the precision reaches approximately 30 to 95. Consequently, the precision value for the proposed method is 0.9412, respectively.

Figure 6 portrays the performance graph for recall, however, the recall of the proposed method decreases with the increasing number of epochs. Subsequently, the recall is decreased approximately 90 to 0, when increasing the epochs like 0 to 100. Consequently, the recall value for the proposed method is 0.873, respectively.

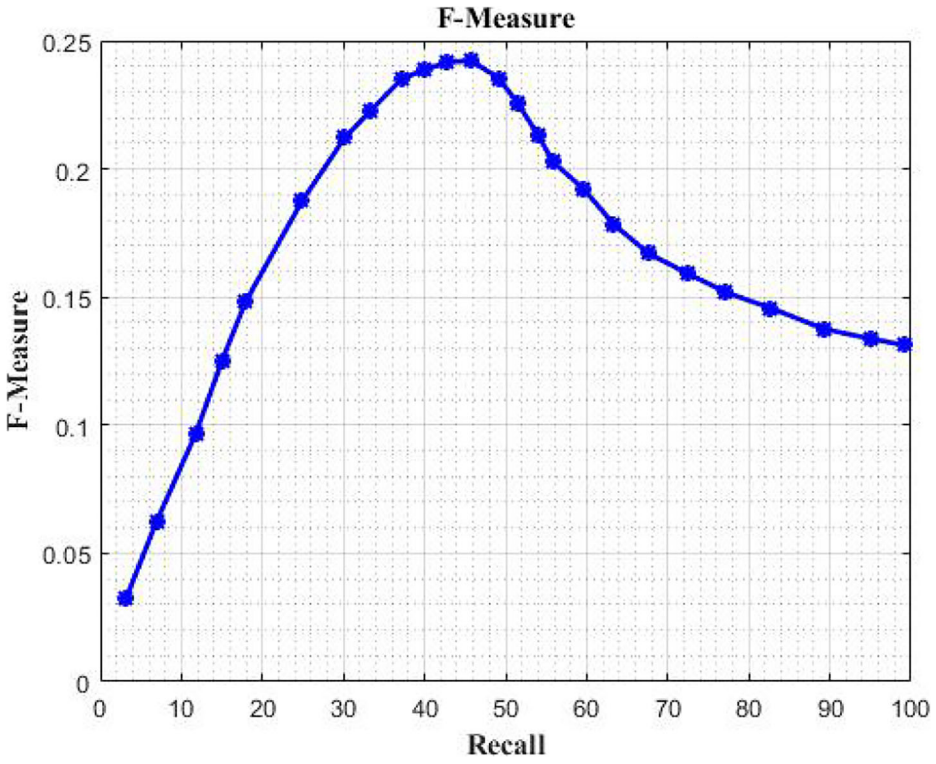


Fig. 7 F-Measure Vs Recall

F-Measure provides a single score that balances both precision and recall issues in a single number. Recall and F- measures are calculated by using the following equations,

$$\text{Re}call\,(R) = \frac{N_k}{N} \tag{23}$$

$$F{-}measure\,(F) = \frac{2 \times P \times R}{P + R} \tag{24}$$

Figure 7 depicts the graph of F-measure vs recall performance analysis. Subsequently, this measure is approximately the average of the two when they are close, and is more generally the harmonic mean, which, for the case of two numbers, coincides with the square of the geometric mean divided by the arithmetic mean.

This paper discusses the implementation of an effective algorithm for video noise. The analysis of the performance of a proposed algorithm is carried out by the peak signal to noise ratio (PSNR) to evaluate the reduction of noise. It is computed between each frame and the video signal degraded.

Figure 8 portrays the performance analysis of the peak signal to noise ratio. The PSNR of the proposed technique is 16 dB. The figure illustrates that the proposed method highly reduced the noise amount with the utilization of the Motion Adaptive Gaussian denoising

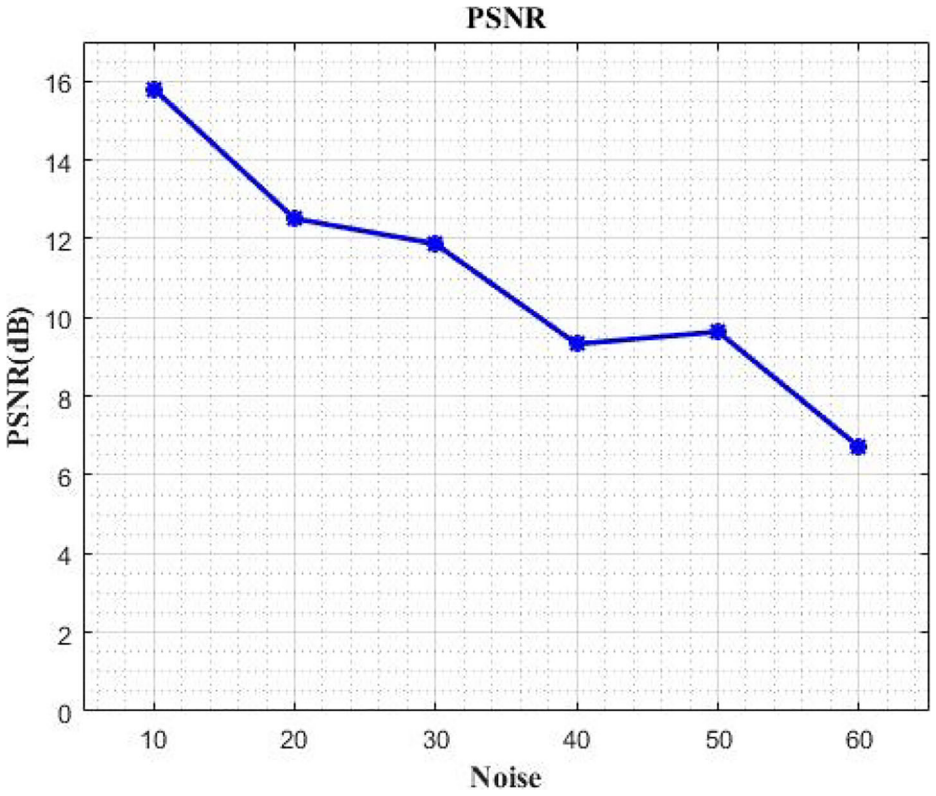


Fig. 8 Peak Signal to Noise Ratio

Filtering method. Consequently, the proposed work achieves high quality with enhanced video.

Figure 9 illustrates the accuracy of the proposed technique which is about 0.94%. The accuracy of validation in several periods is a good way to determine whether the model has been adequately trained. Accordingly, the content-based extraction technique achieves a high level of accuracy in terms of several low-level features.

Figure 10 portrays the NDCG for the optimized feed-forward neural network. Normalized Discounted Cumulative Gain (NDCG) is a common way to measure a search result's efficiency. Consequently, this predicts the result with accurate features and the given input data with high-quality output.

Figure 11 examines the balanced accuracy of the proposed uncertainty estimation technique. These results are calculated in conjunction with entropy (0.95), variance (0.942), and BC (0.7). These findings demonstrate that uncertainty indices may be helpful to classify samples where the classifier is susceptible to correct forecasts by the Monte Carlo method of sampling.

5.3 Comparison analysis

The proposed method's performance is compared with the different existing techniques, that the existing methods are MMM (Multi-source Multi-net Micro-video Recommendation), LP-

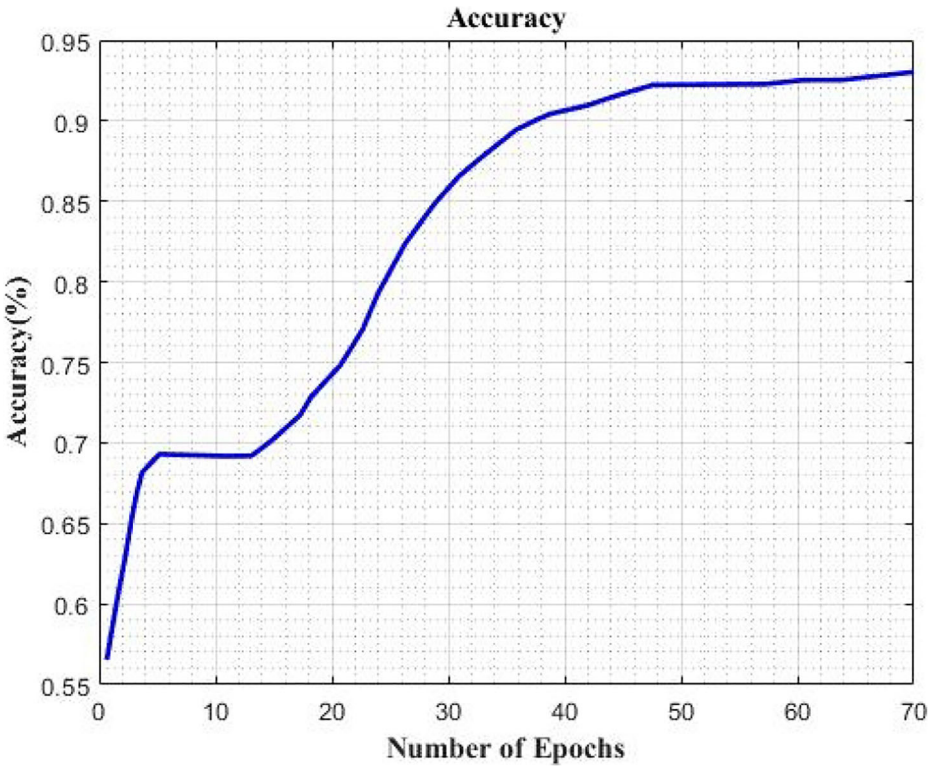


Fig. 9 Accuracy of Proposed Technique

LGSN (Link Prediction Considering Local and Global Structures of the Network), and CDPRec (Context-Dependent Propagating Recommendation network) [12, 13, 18]. The comparison parameters are precision, recall, F- measures, and nDCG, compare these parameters with the existing technique.

Table 3 portrays the comparison table of the proposed method with the existing MMM, CDPRec, and LP-LGSN methods. The proposed method compares the parameters like precision, recall, F-measures, and nDCG with the existing methods. Subsequently, the training percentage is 10, 20, 30, and 40. Consequently, it depicted that the proposed method produces the good values than the other methods, which can be represented in Table 3 and their graphical parts are represented as follows.

Figure 12 elucidate the comparison graph for precision, which compares the proposed method with the existing MMM, LP-LGSN, and CDPRec techniques. Precision is the degree to which two or more measurements are close to each other. The precision value is executed for the training percentage 10, 20, 30, 40 and the precision value for the existing method is 0.03436 for MMM, 0.0386 for CDPRec, and 0.903 for LP-LGSN, compare to these values proposed method value is high at 0.937. The performance of the proposed method is 2% higher than the existing LP-LGSN, MMM, and CDPRec techniques.

Figure 13 depicts the comparison graph for recall with the existing techniques MMM, LP-LGSN, and CDPRec. To find all the positive samples, recall is intuitively the ability of the

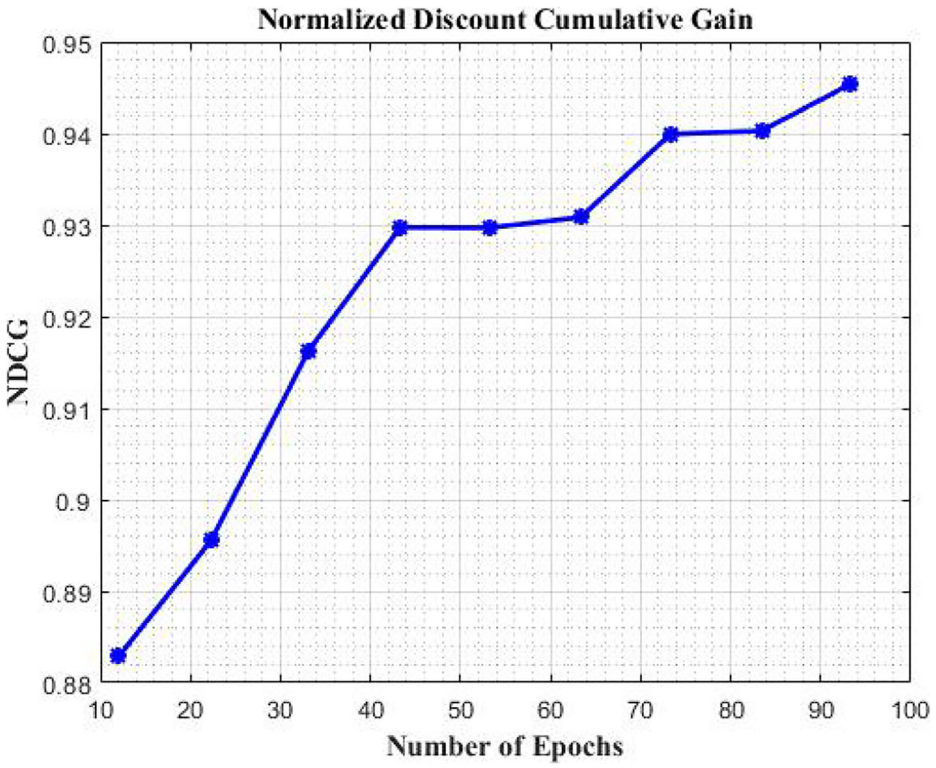


Fig. 10 Optimizing NDCG using Feedforward Neural Network

classifier. The recall is executed at the training percentages 10, 20, 30, and 40. The existing methods recall value is 0.400 for MMM, 0.0307 for CDPRec, 0.090 for LP-LGSN, when compared to these values proposed value is obtaining as high, i.e., 0.5176. Therefore, the performance of the proposed method is 3% higher than the existing CDPRec, LP-LGSN, and MMM methods.

Figure 14 exemplifies the performance comparison of F-measures; it compares the proposed method with the existing LP-LGSN and CDPRec. A single score is generated by F-Measure that balances both precision and recall concerns in a single number. The F-Measure value is executed due to the training percentage of 10, 20, 30, and 40. The existing technique values are 0.0341 for CDPRec and 0.164 for LP-LGSN, when compared to these methods the proposed technique value is obtained as high as that is 0.265. The performance of the proposed method is 5% higher than the existing LP-LGSN and CDPRec methods.

Figure 15 exemplifies the comparison graph of nDCG, thus the nDCG of the proposed technique is compared with the existing MMM method. The quality of a set of search results is assessed by the popular method called Normalized Discounted Cumulative Gain (NDCG). The existing value is executed as 0.32613 for MMM, compare to this technique the value obtained for the proposed technique is high as 0.41863 and the training percentages are 10, 20, 30, 40. The proposed method's performance is 7% higher than the existing MMM method.

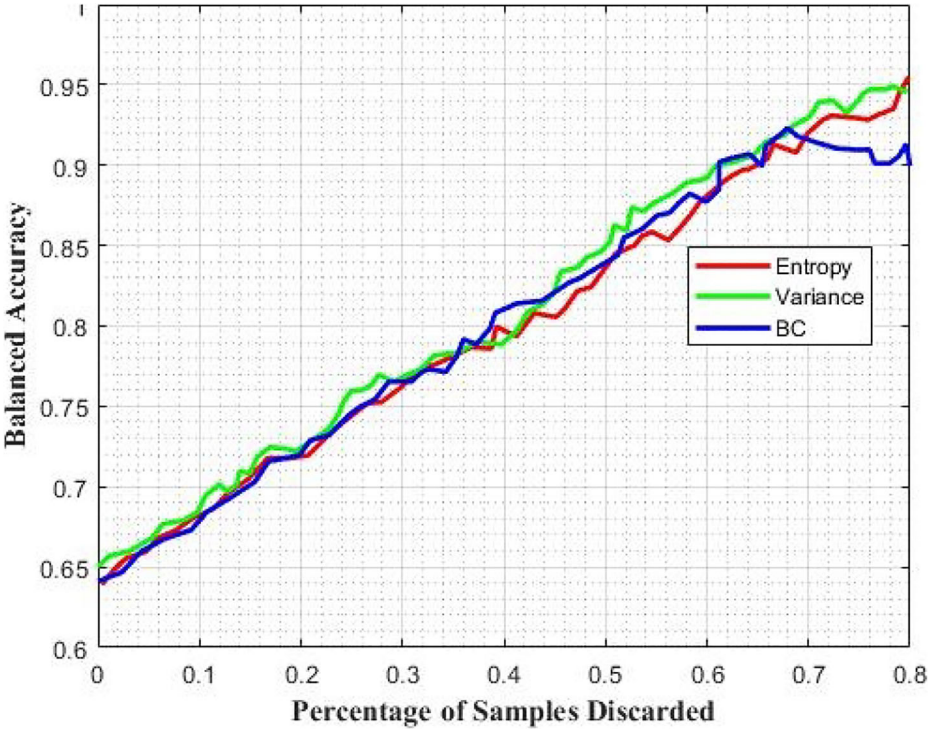


Fig. 11 Balanced Accuracy

6 Research conclusion

Video recommendation system provides users with suitable videos for users to choose from, which is an effective way to get higher user satisfaction. The capability of filtering large information spaces and selecting the items that are likely to be more

Table 3 Comparison Table of Proposed Method for Prediction Parameters

Methods		MMM	CDPRec	LP-LGSN	Proposed Method
Precision	10%	0.1034	0.0357	0.660	0.692
	20%	0.06102	0.0374	0.808	0.876
	30%	0.04374	0.0401	0.874	0.904
	40%	0.03436	0.0386	0.903	0.937
Recall	10%	0.3257	0.0289	0.044	0.389
	20%	0.365	0.0291	0.061	0.435
	30%	0.385	0.0314	0.075	0.4698
	40%	0.400	0.0307	0.090	0.5176
F-Measures	10%	–	0.0319	0.082	0.102
	20%	–	0.0327	0.113	0.159
	30%	–	0.0352	0.137	0.247
	40%	–	0.0341	0.164	0.265
nDCG	10%	0.29904	–	–	0.3795
	20%	0.31502	–	–	0.37269
	30%	0.32192	–	–	0.39470
	40%	0.32613	–	–	0.41863

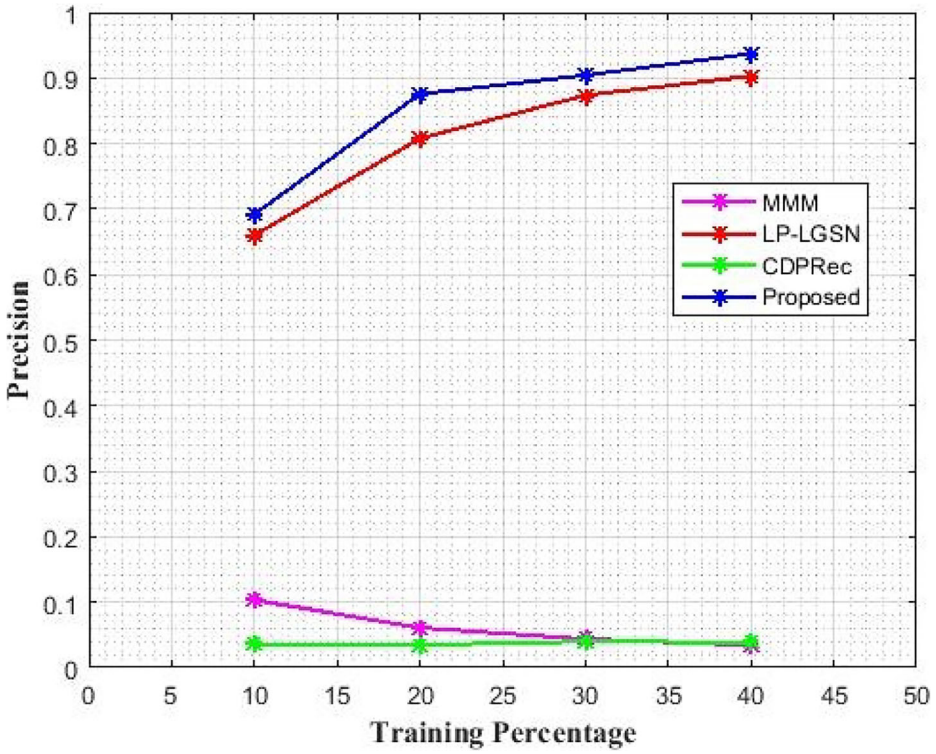


Fig. 12 Precision Comparison Graph

attractive and interesting to a user is characterized by the Recommendation Systems (RSs). A Feedforward neural network-based Monte Carlo sampling technique is proposed to predict the video based on their visual features. Initially, unwanted noises from the videos are removed by using the Motion Adaptive Gaussian denoising Filtering and it gives a high-quality video and clarity video frame. A content-based extraction technique is used to extract the visual features of the video. The visual features like object motion, lightning key, shot duration, and colour variance are extracted from these videos. Finally, these extracted features and user input are given to the Multilayer feedforward neural network and this predicts the relevant video. To avoid uncertainty problems, the work incorporates the Monte Carlo sampling technique, which predicts the accurate video according to the recommendation.

The videos are collected from the standard UGC dataset. In this research 170 videos are collected for prediction by using the proposed technique. This was implemented and executed in the MATLAB r2020a software. The precision, recall, F-measures, and nDCG are the comparison parameters of this research. The performance of this proposed method is 2% higher than the existing MMM (Multi-source Multi-net Micro-video Recommendation), LP-LGSN (Link Prediction Considering Local and Global Structures of the Network), and CDPRec (Context-Dependent Propagating Recommendation network) methods for precision. The performance of the proposed method is 3% higher than the existing MMM, LP-LGSN, and CDPRec methods and 5% higher than LP-LGSN, and CDPRec techniques. Additionally, for nDCG, the performance of the proposed method is 7% higher than

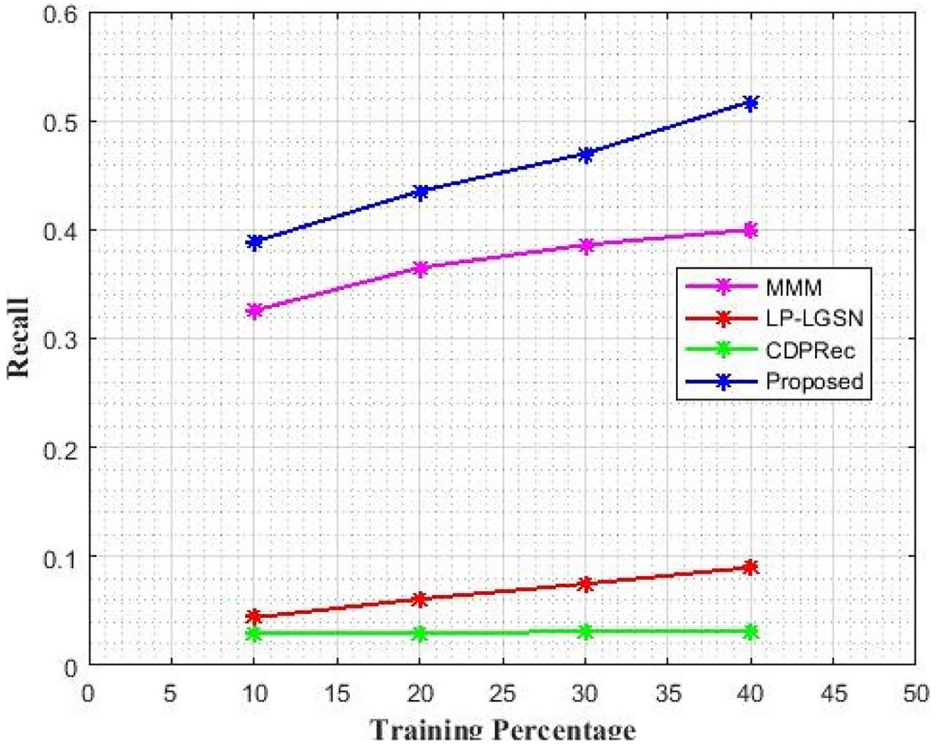


Fig. 13 Recall Comparison Graph

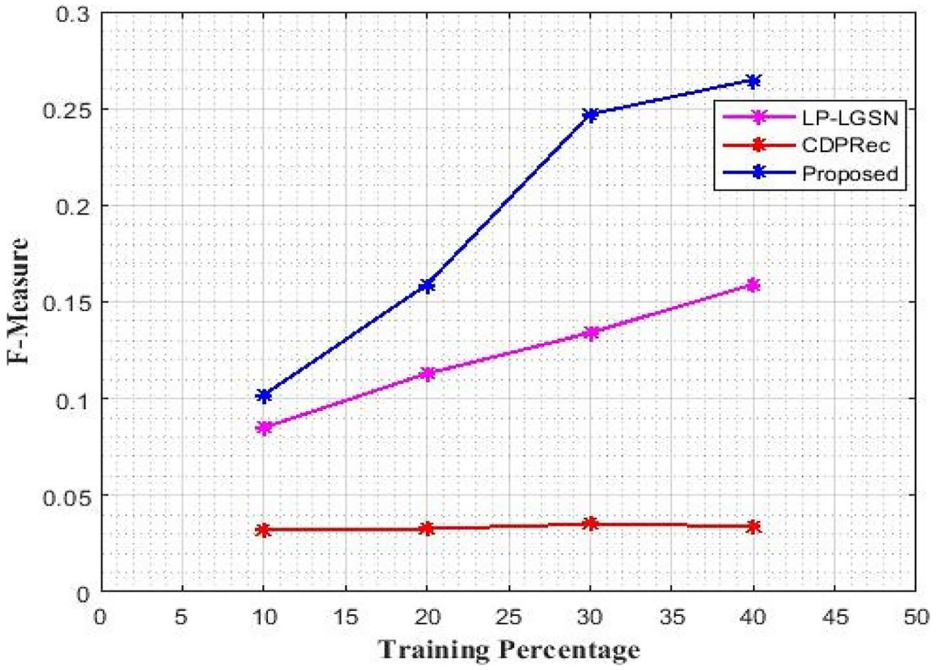


Fig. 14 F-Measures Comparison Graph

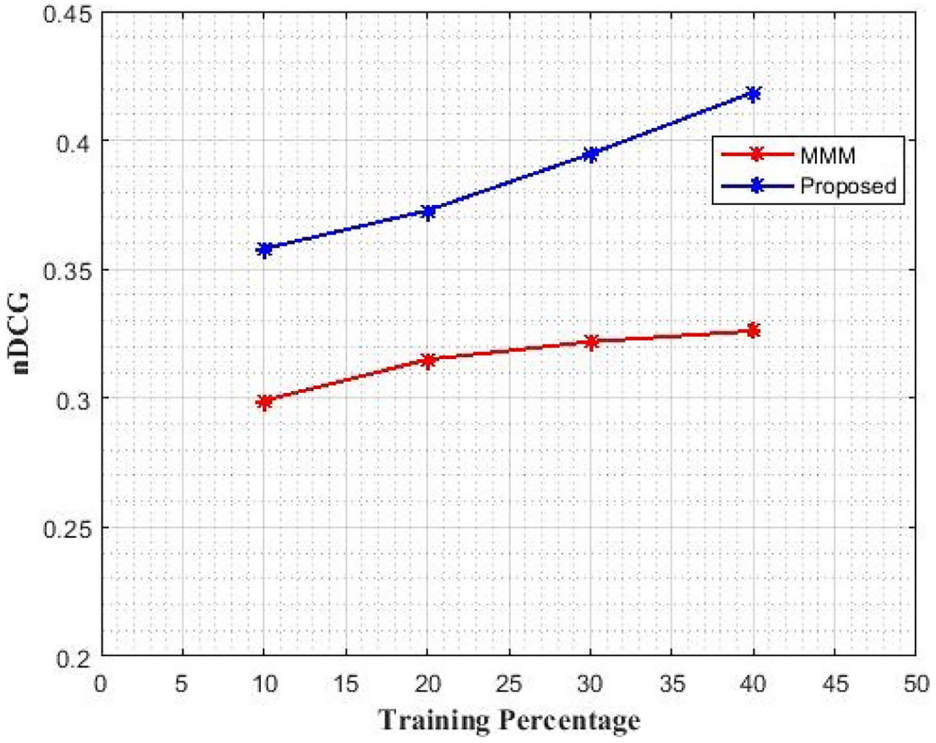


Fig. 15 Comparison of nDCG

the existing method. Subsequently, the proposed method has a less execution time of 0.999 s, furthermore, the results of the proposed method outperform other existing methods, and the proposed method produces higher accuracy of 0.94%. Consequently, a novel deep learning-based method can be introduced in future work to accurately predict the user's interested video, and to improve the prediction accuracy, respectively.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Almeida A, et al (2020) Exploring the multimodal information from video content using deep learning features of appearance, audio and action for video recommendation. arXiv preprint arXiv:2011.10834.
2. Álvarez F, Sánchez F, Hernández-Peñaloza G, Jiménez D, Menéndez JM, Cisneros G (2019) On the influence of low-level visual features in film classification. PLoS One 14(2):0211406
3. Choi I, Oh M, Kim J, Ryu Y (2016) Collaborative filtering with facial expressions for online video recommendation, international journal of information management, 36(3):397–402Available. https://doi.org/10.1016/j.ijinfomgt.2016.01

4. Deldjoo Y, Elahi M, Quadrana M, Cremonesi P (2018) Using visual features based on MPEG-7 and deep learning for movie recommendation, Int J Multi Inf Retriev, 7(4):207–219 Available. https://doi.org/10.1007/s13735-018-0155-1
5. Du X, Yin H, Chen L, Wang Y, Yang Y, Zhou X (2020) Personalized video recommendation using rich contents from videos, IEEE Trans Knowledge Data Eng, 32(3):492–505 Available. https://doi.org/10.1109/tkde.2018.2885520
6. Duan S, Zhang D, Wang Y, Li L, Zhang Y (2020) JointRec: a deep-learning-based joint cloud video recommendation framework for Mobile IoT, IEEE Int Things J, 7(3):1655–1666 Available. https://doi.org/10.1109/jiot.2019.2944889
7. Hazrati N, Elahi M (2020) Addressing the new item problem in video recommender systems by incorporation of visual features with restricted Boltzmann machines. Exp Syst Available. https://doi.org/10.1111/exsy.12645
8. Kaklauskas A et al (2018) A neuro-advertising property video recommendation system, Technol Forecast Social Change, 131:78–93Available. https://doi.org/10.1016/j.techfore.2017.07.011
9. Khan A, Shao J, Ali W, Tumrani S (2020) Content-aware summarization of broadcast sports videos: an audio–visual feature extraction approach, Neural Processg Lett, 52(3):1945–1968 Available. https://doi.org/10.1007/s11063-020-10200-3
10. Ling S, Baveye Y, Le Callet P, Skinner J and Katsavounidis I (2020) July. Towards perceptually-optimized compression of user generated content (ugc): Prediction of ugc rate-distortion category. In 2020 IEEE international conference on multimedia and expo (ICME) 1–6. IEEE.
11. Ma J, Li G, Zhong M, Zhao X, Zhu L, Li X (2017) LGA: latent genre aware micro-video recommendation on social media, Multi Tools Appl, 77(3):2991–3008 Available. https://doi.org/10.1007/s11042-017-4827-2
12. Ma J, Wen J, Zhong M, Chen W, Li X (2019) MMM: multi-source multi-net micro-video recommendation with clustered hidden item representation learning, Data Sci Eng, 4(3):240–253 Available. https://doi.org/10.1007/s41019-019-00101-4
13. Matsumoto Y, Harakawa R, Ogawa T, Haseyama M (2019) Music video recommendation based on link prediction considering local and global structures of a network, IEEE Access, 7:104155–104167 Available. https://doi.org/10.1109/access.2019.2930713
14. Mehta (2017) Sandip. Gaussian Noise Removal Using Multiple Wavelets Approach 9(1):61–66
15. Pu S et al (2020) Multimodal Topic Learning for Video Recommendation. arXiv preprint arXiv: 2010.13373.
16. Sajib MSR et al (2018) Video recommendation system for YouTube considering users feedback. Global J Comput Sci Technol 18(1)
17. Sang L, Xu M, Qian S, Martin M, Li P, Wu X (2020) Context-dependent propagating based video recommendation in multimodal heterogeneous information networks. IEEE Trans Multimedia
18. Sang L, Xu M, Qian S, Martin M, Li P, Wu X (2020) Context-dependent propagating based video recommendation in multimodal heterogeneous information networks. IEEE Trans Multimedia. 1-1. Available. https://doi.org/10.1109/tmm.2020.3007330
19. Sun L, Wang X, Wang Z, Zhao H, Zhu W (2017) Social-aware video recommendation for online social groups, IEEE Trans Multimedia, 19(3):609–618 Available. https://doi.org/10.1109/tmm.2016.2635589
20. Tahmasebi H, Ravanmehr R, Mohamadrezaei R (2020) Social movie recommender system based on deep autoencoder network using twitter data. Neural Comput Appl Available. https://doi.org/10.1007/s00521-020-05085-1
21. Tippaya S, Sitjongsataporn S, Tan T, Khan M, Chamnongthai K (2017) Multi-modal visual features-based video shot boundary detection, IEEE Access, 5:12563–12575 Available. https://doi.org/10.1109/access.2017.2717998
22. Tripathi A, Ashwin T, Guddeti R (2019) EmoWare: a context-aware framework for personalized video recommendation using affective video sequences, IEEE Access, 7:51185–51200 Available. https://doi.org/10.1109/access.2019.2911235
23. Wang X, Gao C, Ding J, Li Y, Jin D (2019) CMBPR: category-aided Multi-Channel Bayesian personalized ranking for short video recommendation. IEEE Access. 7:48209–48223 Available. https://doi.org/10.1109/access.2019.2907494
24. Wua T, Lib Y, Wangc Y (n.d.) Personalized recommendation system of UGC (User Generated Content) video resources based on user interest graphs. Acad J Comput Inf Sci 3(1):142–149
25. Xanat V, Toshimasa Y (2019) A video recommendation system for complex topic learning based on a sustainable design approach, Vietnam jJ Comput Sci, 06(03):329–342 Available. https://doi.org/10.1142/s2196888819500179
26. Yan H, Yang C, Yu D, Li Y, Jin D, Chiu D (2021) Multi-site user behavior modeling and its application in video recommendation, IEEE Trans Knowl Data Eng, 33(1):180–193Available. https://doi.org/10.1109/tkde.2019.2926078

27. Zhou X et al (2017) Enhancing online video recommendation using social user interactions. VLDB J, 26(5): 637–656 Available. https://doi.org/10.1007/s00778-017-0469-2