Journal of Applied Research and Technology

www.jart.icat.unam.mx

ICAT Instituto de Ciencias Aplicadas y Tecnología

Journal of Applied Research and Technology 23 (2025) 463-479

Original

Design of an iterative model for educational video classification using graph-based self-training methods

M. Choudhary[1]* • S. Rungta[1] • S. Pandey[2]

[1]*Department of Computer Science and Engineering, RCET, Bhilai, Chhattisgarh, India*
[2]*Department of Computer Science and Engineering, BIT, Durg, Chhattisgarh, India*

Received 11 04 2024; accepted 01 15 2025
Available 10 31 2025

Abstract: The critical need for advanced approaches in classifying educational videos has been extensively researched, as the domain has faced significant challenges due to the limited availability of labeled data, noisy annotations, and the inherent diversity of video content. Traditional approaches are likely to fall short in managing such complexity, resulting in suboptimal classification performance. A new integration of graph-based semi-supervised learning (GSSL), self-training with consistency regularization, adversarial learning, transfer learning from pretrained models, and a weakly supervised learning framework is proposed in this paper. All these approaches help improve performance in our proposed framework across several metrics, including increased precision, accuracy, recall, and the area under the curve (AUC), which subsequently reduces delays and increases specificity with respect to the EDUVSUM and HowTo100M datasets. The uniqueness of combining these techniques will also enhance the classification accuracy and competence of such models, resulting in a robust and generalizable classifier across various domains of educational content. This paper presents a significant contribution to the field of educational video classification by providing a comprehensive solution to the multifaceted challenges of the task.

Keywords: **Educational Video Classification, Graph-Based Semi-Supervised Learning, Adversarial Learning, Transfer Learning, Weakly Supervised Learning, Process**

*Corresponding author.
E-mail address: minukum@gmail.com (M. Choudhary).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

1. Introduction

With the soaring use of digital educational content, especially videos, there is a need for sophisticated classification systems to facilitate the effective organization, search, and recommendation of these resources to learners. The conditions under which educational videos are considered are unique and may not entirely be handled by conventional video classification frameworks. The reasons include the limited availability of labeled data, the presence of noisy annotations, and the diverse range of video content, which spans lectures to interactive tutorials. It complicates the task, mandating a more nuanced approach to the educational video classification process (Kadam et al., 2022; Wang et al., 2021; Apostolidis et al., 2021)

While traditional classification methodologies have made remarkable strides in addressing some of these challenges, they often fall short when confronted with the complex and multifaceted nature of educational content. The lack of labels in existing models is primarily attributed to the reliance on extensive labeled datasets and samples (Zhang et al., 2023; Davila et al., 2021; Yuan & Zhang, 2023), susceptibility to annotation noise, and the inability to generalize across a wide range of educational topics and presentation styles. The current work addresses this limitation by proposing a comprehensive framework that provides a holistic solution to transcend traditional models, adopting a synergistic integration of modern machine learning techniques.

This paper presents a new framework that combines GSSL, self-training with consistency regularization, adversarial learning, transfer learning from pre-trained models, and weakly supervised learning. This was considered with the view of drawing from these methodologies, as they already exhibit improved efficacy in circumventing some of the challenges posed by the classification of educational videos. This relies on GSSL, whereby the relational structure of the data is utilized to facilitate effective label propagation from a limited number of labeled samples to a larger set of unlabeled ones, thereby mitigating the problem of scarce labeled data samples.

Self-training with consistency regularization has a foolproof mechanism that enables the improvement of model reliability, particularly in resolving ambiguous or noisy data by ensuring predictive consistency across various adversarial perturbations of input data samples. Adversarial learning is applied to teach invariance with respect to the domain since, without it, the model would not be able to generalize over the broad spectrum of educational content. Transfer learning has integrated transferred models that learn from vast, diverse datasets, significantly reducing the dependence on labeled data from a particular domain. Lastly, weakly supervised learning techniques can utilize imperfectly labeled data to generate meaningful patterns that the model can then leverage.

A combination of these advanced techniques (Liu et al., 2022; Nagar et al., 2021; Ma et al., 2022) within a united framework has represented a giant leap forward in the era of digital educational video classification. This approach not only addresses the challenges mentioned above but also lays down a benchmark for classification accuracy, robustness, and adaptability. The proposed framework enables the easy and effective use of both labeled and unlabeled data, thereby making the model robust against noisy annotations and providing a generalizing model that performs well across diverse educational materials. This makes the framework a key contribution to the domain of digital learning resources.

This paper, as an introduction, will detail the selected methodologies and their complementary strengths, setting the stage for a comprehensive exploration of the proposed process. In the following sections, it will delve into the technical intricacies of each component, their integration, and the empirical evaluation of the framework, showcasing its superior performance compared with the existing methodologies and underscoring its potential to transform the landscape of educational video classification processes.

Motivation & Contribution

The availability of vast amounts of digital educational resources, based on video, has prompted an urgent call for sophisticated classification systems. Such systems are important since they help streamline users' searching, interpreting, and selecting learning materials. Nevertheless, conventional classification strategies face formidable barriers due to the intrinsic complexities of educational video content, such as diverse instructional styles, varied subject matter, and fluctuating quality. The scarcity of labeled data, the presence of noisy annotations, and the dynamic nature of educational content all complicate the challenges, making it difficult to implement conventional classification methods properly. Attempted in this study, therefore, is that of banning such methodologies by making use of advanced machine learning techniques available in other and similar domains, but yet untapped in the domain of educational content.

The scope of the proposals is marked by several contributions, reflecting a fresh research contribution in the

realm of educational technology. First and foremost, the proposed framework comprises an integrative fusion of GSSL, self-training with consistency regularization, adversarial learning, transfer learning from pretrained models, and weakly supervised learning. It integrates these and is strongly designed to gain full benefits from the effective strengths of each method in a holistic manner towards the resolution of the classifier challenges at hand related to educational video classification. The imposition of GSSL, for example, builds up the data-propagating power of the model efficiently for the learning platform and suitably exploits the intrinsic structure of the data samples. Concurrently, the application of adversarial learning techniques promotes the development of domain-invariant features, enabling robust classification across diverse educational content.

Secondly, this framework employs a strategic approach to utilizing both labeled and unlabeled data, thereby mitigating the limitations imposed by the scarcity of annotations. Self-training with consistency regularization is employed to achieve stable predictions, even with ambiguous data, thereby enhancing the model's generalizability and improving its overall performance. Besides, by transferring learning from pretrained models, the incorporation indicates the versatility of such a framework for deriving and applying previously learned knowledge bases in the absence of extensive domain-specific labeled datasets and samples.

Finally, the use of weakly supervised learning addresses the widespread issue of noisy and incomplete annotations. Since the model is robustly designed to handle spurious annotations, learn from available data, and, importantly, mitigate the effects of imprecise labels, learning from data is done effectively, even with such high quality and quantity.

These factors translate to an inclusive and sturdy model to set new benchmarks in the classification of educational videos. The proposed model has successfully outperformed existing methodologies on benchmark datasets such as EDUVSUM and HowTo100M, using 3 or more superior performance metrics in comparison. This work addresses the immediate tasks that educators, learners, and technologists face in the digital age, while also leaving numerous opportunities for future enhancement of such technologies.

2. In-depth review of existing Models

Digital education and the rapid proliferation of video content require further advancements in their classification and summarization to enhance accessibility and learning efficiency. The standard approaches to video summarization have been developed around general content, ignoring most cases and the diversified requirements of educational videos and samples. This gap in research motivates the exploration of novel frameworks and methodologies for educational content that can address challenges such as limited labeled data, noisy annotations, and diverse content levels (Issa & Shanableh, 2022; Mujtaba et al., 2022).

In recent times, several new methods related to machine learning have been introduced, ranging from deep neural networks to graph-based approaches, to enhance video summarization. These methods have proven effective in various contexts, including medical education, personal video summarization, and unsupervised learning (Ji et al., 2021; Apostolidis et al., 2021). However, the methods above are not entirely applicable when it comes to classifying and summarizing educational videos. There is a need for domain-specific adaptation, as well as the management of specific academic content characteristics.

The recent research on video summarization, as presented in the current literature review, has identified one of the many approaches, each with its own strengths and weaknesses. Methods such as Motion-Assisted Reconstruction Network (MAR-Net) and Deep Reinforcement Learning with Shot-Level Semantics demonstrate promising results in capturing dynamic content. Still, both will most likely be required for placing in educational narratives rather than for motion or scene changes (Ma et al., 2020a; Gao et al., 2021). Such a feature, focusing on extractive summarization for lecture videos, signifies direct applications in educational content, which underline the effectiveness of domain-specific approaches. This is because it puts heavier emphasis on whiteboard or chalkboard content. Therefore, it fails to capture the variety within educational videos, such as interactive tutorials or practical demonstrations. Relational reasoning, particularly in the context of spatial-temporal graphs and affective visual information for summarization, highlights the potential to leverage complex data representations and human-centric cues. Such approaches suggest that a comprehensive framework for educational video summarization could be enhanced by using multimodal data and emotional engagement metrics to better align with educational outcomes (Zhao et al., 2021a; Zhang et al., 2022). Methods that utilize unsupervised learning are ways to overcome the limitations of labeled data in the educational learning context. However, implementing such

Table 1. Different methodologies based on the literature review

Sl. No.	Method Used	Findings	Results	Limitations
1	Machine Learning Algorithms for Video Summarization	Explored challenges and opportunities in video summarization, emphasizing big data's role.	Highlighted the efficacy of SVS and MVS in handling large datasets.	Limited discussion on the adaptability to educational content.
2	Improved Clustering and Silhouette Coefficient for Keyframe Generation	Proposed an enhanced clustering method for video summarization.	Achieved improved precision in keyframe selection.	Focused mainly on static scenes; may not generalize well to dynamic educational videos.
3	Deep Neural Networks Survey	Reviewed deep learning approaches for video summarization.	Identified the gap between supervised and unsupervised learning techniques.	Lacked specific solutions for educational video content.
4	MAR-Net: Motion-Assisted Reconstruction Network	Utilized motion information and an attention mechanism for summarization.	Demonstrated semantic consistency in unsupervised settings.	A motion-based approach may not fully capture the nuances of educational video content.
5	FCN-LectureNet for Lecture Video Summarization	Focused on extractive summarization of educational content.	Improved detection and summarization of lecture videos.	Primarily targeted at whiteboard/chalkboard content, it may not be comprehensive.
6	Deep Reinforcement Learning with Shot-Level Semantics	Introduced an unsupervised learning model focusing on shot-level semantics.	Reported advancements in summarization without extensive labeling.	Shot-level focus might overlook the broader educational context.
7	3D Spatio-Temporal U-Net via Reinforcement Learning	Applied 3D convolutional networks for medical video summarization.	Showcased potential in medical education videos.	Specificity to medical videos may limit applicability to general education.
8	Personalized Summaries of Egocentric Videos	Developed a reinforcement learning approach for personalization.	Achieved personalized summarization in first-person videos.	Focused on egocentric videos, which represent a niche area within educational content.
9	Adaptive Multiview Graph Difference Analysis	Proposed a novel graph-based method for summarization.	Enhanced adaptability and efficiency in processing.	The complexity of the method may hinder its application in real-time scenarios.
10	CNN and HEVC Features for Static Summarization	Leveraged deep learning and video coding features for summarization.	Improved efficiency in static video summarization.	Limited by its focus on static summarization, it overlooks dynamic educational content.
11	LTC-SUM: 2D CNN for Personalized Summarization	Introduced a lightweight framework using client-driven 2D CNN.	Facilitated personalized video summarization efficiently.	The client-driven approach might not fully address the diversity of educational videos.
12	Deep Attentive Video Summarization with Distribution Consistency	Employed attention mechanisms and consistency learning.	Showed improvement in capturing key video segments.	The method's reliance on deep learning might limit its accessibility for resource-constrained environments.

Sl. No.	Method Used	Findings	Results	Limitations
13	AC-SUM-GAN: Actor-Critic with GANs	Combined GANs with reinforcement learning for unsupervised learning.	Enhanced creative aspects of video summarization.	The complexity and computational demands of GANs may not be suitable for all educational applications.
14	Keyframe Extraction via Dictionary Selection	Applied a dictionary selection approach for keyframe extraction in laparoscopic videos.	Offered a novel solution for medical video summarization.	The focus on laparoscopic videos limits generalization to other educational areas.
15	Relation-Aware Assignment Learning	Introduced an unsupervised approach using graph neural networks.	Advance the efficiency of video summarization.	It may not specifically address the unique challenges of educational video summarization.
16	Audio Visual Video Summarization	Explored multimodal learning for summarization.	Demonstrated the importance of audiovisual cues.	The reliance on multimodal inputs may not apply to all educational videos.
17	Joint Reinforcement and Contrastive Learning	Utilized a novel combination of learning techniques for summarization.	Showed potential in unsupervised learning contexts.	The specific learning approach might complicate implementation.
18	Affective Visual Information for Summarization	Investigated the role of emotion in video summarization.	Highlighted the value of affective cues in human-centric videos.	The focus on affective information may overlook educational content's instructional aspect.
19	Multimodal and Aesthetic-Guided Narrative Summarization	Combined multimodal information with aesthetic guidance.	Enhanced narrative video summarization.	The emphasis on aesthetics might not align with educational video summarization priorities.
20	Similarity-Based Sparse Subset Selection	Developed a kernel sparse representation method for summarization.	Improved the selection of informative video segments.	The kernel approach's complexity might challenge its broader application.
21	Sequence-Graph Network for Summarization	Introduced a reconstructive network for key-shot summarization.	Offered advancements in summary generation.	The focus on key-shot generation might not capture the full educational narrative.
22	TTH-RNN for Video Summarization	Applied tensor-train hierarchical RNNs for efficient summarization.	Demonstrated the potential of hierarchical structures.	The specialized network architecture may limit its adaptability.
23	EEG-Video Emotion-Based Summarization	Explored EEG signals for emotion-based summarization.	Provided insights into multimodal emotion recognition.	The niche focus on EEG-video data may not be universally applicable.
24	CoEvo-Net for Highlight Detection	Developed a coevolution network for video analysis.	Addressed the effective detection of video highlights.	The specific focus on highlights may miss comprehensive educational content.
25	Spatial-Temporal Graphs for Summarization	Utilized relational reasoning over spatial-temporal graphs.	Showed improvements in summarization through graph-based techniques.	The method's focus on spatial-temporal graphs may not suit all video types.

techniques requires extra care to ensure that the summary, whether on video or audio tape, maintains instructional integrity and relevance, particularly in educational video summarization contexts (Köprü & Ezrin, 2023).

In general, the review highlights the need for a multifaceted approach to educational video summarization, which combines the strengths of existing methods and addresses their limitations within the educational domain (Zhao et al., 2021b). In fact, a robust solution may comprise an adaptive learning platform, feature learning tailored to the specific domain, and a comprehensive process for multimodal data processing that would enable significantly improved classification and summarization of educational video content. Such a comprehensive solution, which effectively addresses the difficulties and lack of user engagement in educational video content, sets a solid foundation for future research in the field by guiding the development of more sophisticated and education-oriented video summarization technologies (Xie et al., 2023; Ma et al, 2020b). According to Zhao et al. (2020), to alleviate the lags and expenses associated with rollouts, a range of video summarization models was presented, categorized by their functionality. Based on this discussion, researchers will be in a position to select models that are optimum for their functionality-based use cases. Authors (Lew et al, 2022) presented a time synchronization module that employs an attention mechanism to map EEG representations into a visual representation space. Authors Chen et al. (2022) presented a new model for VHD termed Coevolution Network (CoEvo-Net), which enables the efficient integration of video and language features through the joint evolution of these features, as the process involves the coevolution of two different features from the two modalities. Such a cell is the CoEvo-Cell structure, which integrates language and video, cross-modulates, and removes specific non-essential components of the input, such as word elements within a sentence. Zhu et al. (2022) proposed a dynamic graph modeling approach to learn spatial-temporal representations for video summarization.

Design of the Proposed Model Process

The section then discusses the design of the proposed model that amplifies the efficiency of the summarization process. In the context of educational video classification, GSSL emerges as a vital approach that leverages the inherent data structure and relationships among video samples to enhance label propagation. The essence of GSSL is grounded in the construction and optimization of a graph where nodes represent video samples and edges signify the relationships or similarities between these samples. Let us now delve into the mathematical formulations and iterative processes that underpin the GSSL mechanism, reflecting its application from the initial input of collected video samples to the eventual output elucidating relationships between these samples. The GSSL process starts with the construction of a similarity graph G = (V, E), where V is the set of nodes corresponding to video samples and E is the set of edges connecting these nodes. The edge weights *Wij* between nodes i and j are determined by a similarity function that utilizes a Gaussian kernel, as expressed in Equation 1.

$$Wij = exp\left(-\frac{\|xi-xj\|^2}{2\sigma^2}\right) \quad (1)$$

Where xi*j* are feature representations of video samples, and σ controls the width of the Gaussian kernel. Upon establishing the graph, the GSSL framework incorporates the label information into the graph via a label matrix *Y*, where *Yil = 1* if sample *i* is labeled with class *l* and 0 otherwise. In the semi-supervised setting, most of the samples are unlabeled, thus requiring the propagation of label information from labeled to unlabeled nodes. This propagation is governed by the label propagation matrix *F*, where each element *Fil* means the probability that node *i* belongs to class *l*. To refine the label propagation, the GSSL framework employs an optimization objective that minimizes the discrepancy between predicted and actual labels for labeled samples while ensuring smoothness in label distribution over the graph. This is articulated through the optimization task represented via equation 2.

$$\min^{FTr}(FTLF) + \mu \parallel F - Y \parallel F^2 \quad (2)$$

Where *L=D−W* is the Laplacian matrix, *D* is the diagonal degree matrix, μ is a regularization parameter, and *F* represents the Frobenius norm levels. The equilibrium of the optimization task is reached when the derivative of the objective function with respect to *F* vanishes in the process. This condition yields the equilibrium equation, where *I* is the identity matrix for the process. The resolution of this equation involves the calculation of F, which is iteratively updated via equation 3.

$$F(t+1) = \alpha SF(t) + (1-\alpha)Y \quad (3)$$

Where $S = D^{-1/2}\ WD^{-1/2}$ is the normalized similarity matrix and α is a factor controlling the trade-off between the original label assignments and the propagated labels.

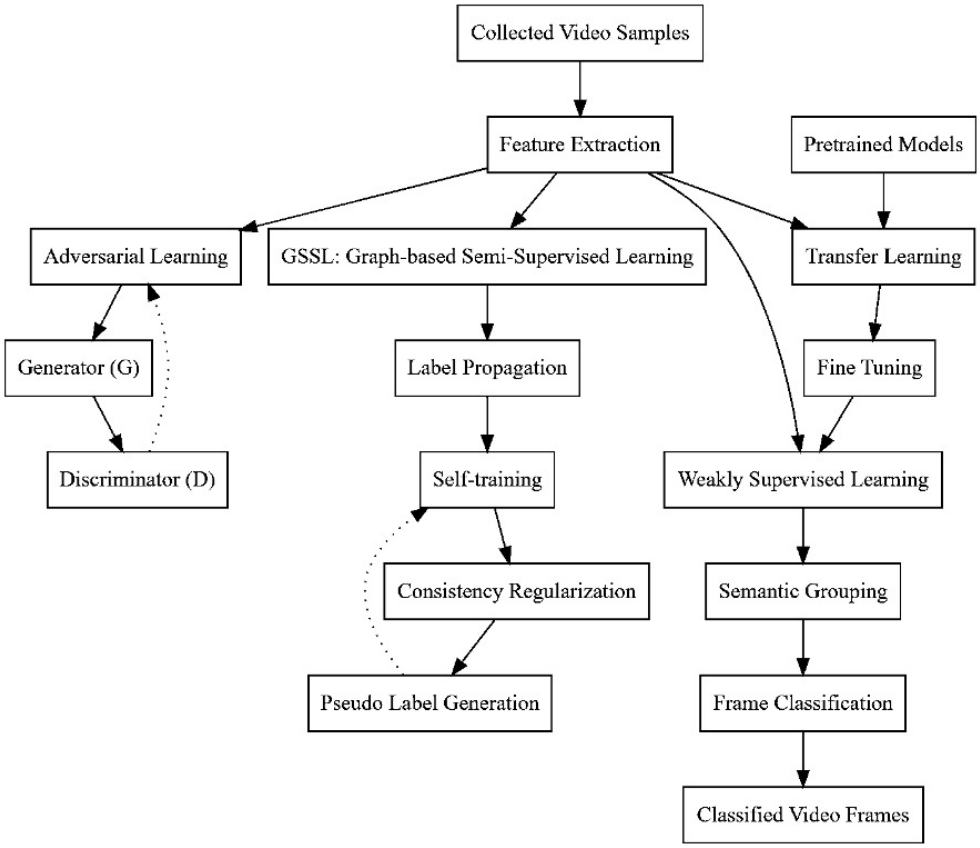

Figure 1. Model Architecture for the Proposed Video Classification Process

As per figure 1, the iterative process continues until convergence, measured by a threshold ϵ, such that $F(t+1)$-$F(t)\|<\epsilon$ for this process. Upon merging, the final F matrix would summarize not only propagated labels but also the inherent structures and similarities amongst the video samples. In this regard, it translates graph theory and semi-supervised learning to a comprehensive framework in the classification of educational content. It enhances the accuracy and efficiency of educational resource classification operations. Through various operational steps, GSSL develops an effective mechanism to classify educational videos, particularly in cases with sparse labels and complex video content. As shown in Figure 2, the process of self-training with consistency regularization works by leveraging relationships between samples to enhance the reliability of these links. This process involves iterative steps, guided by mathematical formulations that refine the predictive model iteratively using unlabeled data alongside labeled instances to derive a more reliable and robust learning outcome. Initiating the process, a set of video samples can be derived such that for each sample, *xi*, and the corresponding relationships derived from the previous phase, encapsulated in the matrix *R*, where *Rij* represents the relationship between videos *i* and *j*. The objective is to exploit these relationships to foster a consistent and reliable mapping, such that the predicted function f: *xi*↦*y*, with y representing the predicted labels or attributes for these video samples.

The self-training component consists of the iterative updating of the predictive model f. First, under the availability of labeled data *L = {(xi,yi)}*, the model f is trained on it. Second, for every unlabeled sample xu from the unlabeled dataset *U*, the model produces a pseudo-label *y' u = f*(*xu*) sets. These pseudo-labels are then integrated into the training process, albeit with a mechanism to control their influence based on their estimated reliability or confidence, as represented in Equation 4.

$$c(xu) = max(f(xu)) \tag{4}$$

Thus, it reflects the maximum predicted probability across potential classes in different use scenarios. The consistency regularization aspect introduces a perturbation δ to each video sample *xu*, creating a perturbed

version $xu'=xu+\delta$ for this process. The core principle here is to enforce that the model's predictions remain stable or consistent when subjected to small perturbations, thus ensuring that the learned relationships are indeed robust against minor variations or noise in the data samples. This is quantitatively expressed through consistency loss, as formulated in Equation 5.

$$Lcons = \sum_{xu \in U} \| f(xu) - f(xu') \|^2 \quad (5)$$

Thus, emphasizing the drive for minimal divergence between the predictions on original and perturbed samples. Simultaneously, the model refines itself by minimizing a composite loss function $=Lsup+\lambda Lcons$, where L*sup* represents the supervised loss computed on the labeled dataset augmented with high-confidence pseudo-labeled samples, and λ is a regularization parameter modulating the impact of the consistency loss. The supervised loss L*sup* is defined via equation 6,

$$Lsup = \sum_{(xi,yi)\in L} \square(f(xi), yi) + \sum_{xu\in U, c(xu)>\tau} \square(f(xu), y'u) \quad (6)$$

With ℓ representing a loss function such as cross-entropy, and τ a confidence threshold dictating the inclusion of pseudo-labels. The iterative refinement process entails the calculation of gradients ∇fL and updating the model parameters according to a chosen optimization algorithm, employing a stochastic gradient descent process. The update rule is represented via equation 7,

$$f(t+1) = f(t) - \eta \nabla fL \quad (7)$$

Where η represents the learning rate for this process. The evolution continues over multiple iterations, with the updated model progressively honed to generate more accurate and consistent predictions. This iterative enhancement is guided by the underlying objective of achieving minimal discrepancy not only between the predicted and actual labels on the labeled dataset but also ensuring that predictions across perturbed and original versions of video samples remain consistent, thereby fostering a robust learning framework capable of handling the ambiguities and uncertainties inherent in educational video content samples.

Next, the implementation of adversarial learning, primarily achieved through the use of Generative Adversarial Networks (GANs), is identified as one of the most effective techniques for extracting domain-invariant features. This strategy addresses the challenge of video distribution. The methodological paradigm is a fundamental interplay between two separate entities: the generator *G* and the discriminator *D*. Generator *G* and discriminator *D* engage in an adversarial process until a generalization of the features yields domain-agnostic representations. In practice, let us first assume the consistent relationships of the video samples as the input, described in a feature space *X* drawn from the previous stages of the classification framework. Within the *GAN* framework, the objective at this point is to map those features into a new space where domain-specific characteristics are minimized, thereby enabling a more generalized and robust classification capability.

The generator *G*, parameterized by weights θg, turns features *x* from the input video into features that are indistinguishable from real features pertaining to the domain in question, preserving domain-invariant attributes. On the other hand, the discriminator *D*, parameterized

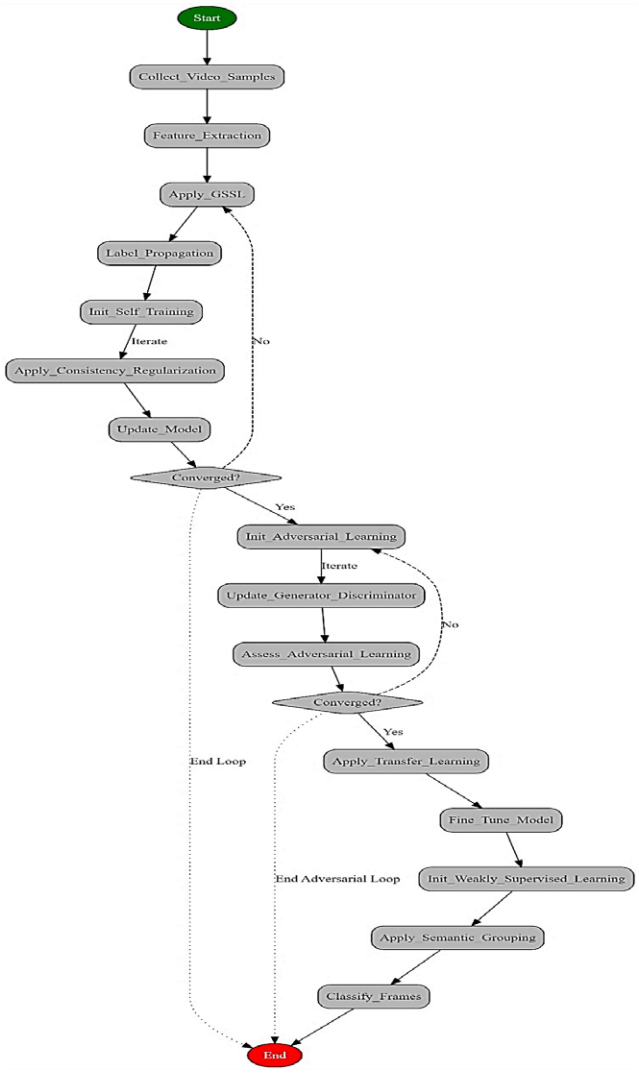

Figure 2. Overall flow of the proposed classification process.

by weights θd, attempts to distinguish between the transformed features generated by G and the features originating from the target domain, essentially evaluating the authenticity of the generated representations. The adversarial learning process is governed by the following min-max game between G and D, encapsulated via equation 8,

$$minGmaxDV(D,G) = Ex \sim pdata(x)[logD(x)] + Ez \sim pz(z)[log(1 - D(G(z)))] \quad (8)$$

Where x represents real features from the target domain, z represents input features or noise variables, *pdata* is the data distribution of real features, and *pz* is the distribution of the input to the generator process. The discriminator D is trained to maximize $V(D,G)$ for a given G, thereby enhancing its ability to distinguish between real and generated features. This is achieved through updating θd by ascending the gradient via equation 9,

$$\nabla \theta d \frac{1}{m} \sum [logD(x(i)) + log(1 - D(G(z(i))))] \quad (9)$$

Where m represents the batch size, and $x(i)$, $z(i)$ are samples from the real data and input distributions, respectively. Simultaneously, the generator G is trained to minimize $V(D,G)$ for a fixed D, aiming to generate features that D will misclassify as real. This is accomplished by updating θg by descending the gradient via equation 10,

$$\nabla \theta g \frac{1}{m} \sum log(1 - D(G(z(i)))) \quad (10)$$

As the adversarial training progresses, G becomes increasingly proficient in creating features that are indistinguishable from real, domain-invariant features, leading to a scenario where D is challenged to differentiate between real and generated samples, symbolizing the achievement of a Nash equilibrium in this adversarial game process.

Then, the results obtained through this process are analyzed using Transfer Learning, which becomes very essential for educational video classification, especially when pre-existing large-scale datasets are utilized to enhance the representation and semantic understanding of extracted features. This approach can bridge limitations due to data scarcity and specificity in targeted educational content by importing and refining knowledge from extensive, diverse sources. The steps used in the process are as follows: At the initial stage, the extracted features are considered a multi-dimensional matrix X obtained from the adversarial learning stage, where every row represents the feature set of a particular video frame. Transfer learning involves mapping these features to a more sophisticated and semantically rich space, a transformation enabled by a neural network model that was initially pre-trained on a large-scale dataset, such as ImageNet or COCO Samples. This model holds a wealth of visual knowledge, encoded by the parameters θ_{pre} for different use case scenarios. This adaptation process goes on in the form of extracting more high-level features, $Z=f\theta pre(X)$, which presents the function represented by the pre-trained network applied to the educational video features X. This is an operation that translates raw, intermediate, or low-level features into a refined feature space, enriched with the broad semantic understanding that has been learned from the samples of the pre-trained dataset samples.

Then, for more specific educational content, a fine-tuning phase is initiated for this process. This updates the model parameters θ from their initial values θ_{pre} to new values θ_{new}, more aligned with the target domain. The fine-tuning is guided by an objective function L(θ) that includes a loss term that quantifies the difference between the actual and predicted semantic categories of the video frames. The update rule follows the gradient descent paradigm, expressed via equation 11,

$$\theta new = \theta pre - \eta \nabla \theta L(\theta) \quad (11)$$

Where η represents the learning rate for this process. The objective function $L(\theta)$ incorporates the cross-entropy loss between the predicted labels and the true labels of the video frames, along with regularization terms to prevent overfitting. This is mathematically described via equation 12,

$$L(\theta) = -\sum \sum yi * clog(pic(\theta)) + \lambda \parallel \theta \parallel^2 \quad (12)$$

Where yi is the binary indicator of whether class c is the correct classification for observation i, $pic(\theta)$ is the predicted probability that observation i belongs to class c, and λ is the regularization parameter for this process. The fine-tuning proceeds iteratively, with each iteration refining the parameters θ to better accommodate the specifics of the educational content, thereby gradually transitioning the model's knowledge base from the general to the particular. This iterative update is mathematically modeled via equation 13,

$$\theta(t+1) = \theta(t) - \eta \nabla \theta L(\theta(t)) \quad (13)$$

Where *t* indexes the iteration rounds. Upon the conclusion of the fine-tuning process, the updated model *fθnew* is employed to reassess the features *X*, resulting in enhanced representations that are inherently more aligned with the semantic intricacies of educational videos and samples. These enriched features lay the groundwork for the semantic grouping of video frames, which can be achieved through clustering techniques, additional classification layers, and effective grouping of frames based on their semantic content and context in various use cases. This transfer learning process culminates in the semantic grouping of video frames, where it represents the unification of general visual knowledge with domain-specific insights, thereby significantly enhancing the capabilities of educational video classification systems to discern and categorize content with higher accuracy and relevance. Transfer learning enables the transfer of knowledge across domains by fine-tuning pre-trained models, ensuring the application of universal visual understanding to the specialized domain of educational content. This approach improves the identification and grouping of semantically coherent frames within educational videos and samples.

Next, for video classification, weakly supervised learning (WSL) is applied, representing a sophisticated technique that uses incomplete or noisy labels to facilitate the discernment of learning cues from the data samples. This holds special significance in situations where obtaining full data annotations is impractical; therefore, WSL leverages available annotations, albeit scant and implicit, for guidance in the learning process. The basic premise of WSL follows that the objective is formulated by integrating uncertainty and ambiguity inherent in weak labels into the learning framework. Consider the input to be the grouped frames categorized semantically, represented by *{Xi},* where every *Xi* constitutes a cluster of frames with similar semantic features. Associated with each group is a weak label *Yi*, which suggests the dominant class among the frames but does not specify the exact label for each frame.

The learning process will then start by defining a probability distribution *P(Y|X;θ)* over possible labels *Y* for a given group *X*, which is parameterized by *θ*. The distribution reflects the model's estimation of the relevance of each label to the grouped frames, where parameters *θ* are to be learned from the data samples. The objective is to optimize *θ* such that *P(Y|X;θ)* takes on a value close to weak labels *Yi* sets. It utilizes a loss function L(θ), which measures the deviation between the predicted and weak labels. Due to the weak nature of *Yi*, additional constraints or regularization terms are often included to guide the learning process. We include a regularization term *R(θ)* that encourages the model to follow assumptions or prior knowledge about the structure and distribution of labels. The combined objective becomes $L(\theta) + \lambda R(\theta)$, where λ is the balancing parameter for this process. The algorithm iteratively adjusts θ to minimize these combined objectives. The update at each iteration *t* is described by the rule represented via equation 14,

$$\theta(t+1) = \theta(t) - \eta\nabla\theta(L(\theta(t)) + \lambda R(\theta(t))) \qquad (14)$$

Where η is the learning rate, and $\nabla\theta$ represents the gradient with respect to θ sets. To account for the inexactitude of weak labels, we incorporate label propagation operations. Each frame *x* in a group *Xi* is assigned a label based on both the group's weak label and the labels of 'nearby' frames, determined by a cosine similarity measure for real-time scenarios. This is expressed as a soft labeling process, where the label assignment for frame *x* is updated via equation 15,

$$Lx = \alpha S(x, Xi)Yi + (1-\alpha)\sum_{x'\in neighbors(x)} S(x, x')Lx' \qquad (15)$$

This is where *Lx* denotes the soft label for frame x; S(·,·) is a similarity function, and α is a parameter that balances the influence of weak labels and neighborhood labels. The WSL process culminates in a model that, despite the initial imprecision of the labels, has distilled meaningful patterns and relationships within the video frames, thereby rendering an enhanced understanding and classification of the content. Iterative refinement of frame labels and model parameters, validated by a combined model-inferred assumptions and empirical data, culminates in robust classification of video frames, transforming ambiguously labeled groups of frames into distinctively classified entities, each aligned with a specific educational theme or topic. This transformation demonstrates how WSL leverages minimal and noisy supervision to derive significant educational insights, facilitating a nuanced understanding and organization of educational video content sets. The performance of this model was evaluated on various scenarios and compared with existing methods in the subsequent section of the text.

3. Result Analysis

This section outlines the experimental setup that will rigorously evaluate the performance of our model for classifying educational videos. This section describes specific configurations, datasets, and parameters applied throughout the experimental phase.

Datasets:

EDUVSUM Dataset: The dataset comprises diverse educational videos across various subjects and educational levels, with each frame labeled as one of the ten categories of educational content. This heterogeneous dataset, among other things, is characterized by its wide range of video quality, presentation styles, and content. EDUVSUM was reshuffled into a ratio of 70%-15%-15% for training, validation, and test sets.

HowTo100M Dataset: This is a large-scale dataset comprising instructional videos covering a broad spectrum of topics from public platforms. The dataset comprises approximately 1.2 million video clips, each accompanied by text descriptions and categorized under 100 different skills and tasks. In this study, we use a subsample of 200,000 clips, representing equally balanced categories for our model. Similar to EDUVSUM, the data was also partitioned into training (70%), validation (15%), and testing (15%) subsets.

Configuration and parameters:

Feature Extraction: We applied a ResNet-50 pre-trained convolutional neural network (CNN) architecture to the initial feature extraction of the video frames. The input to the network consists of standardized video frames, resized to 224x224 pixels.

Graph-Based Semi-Supervised Learning (GSSL): We constructed a graph with video samples as nodes and utilized a k-nearest neighbor algorithm (k = 5) based on cosine similarity for feature categorization, grouping them into skills and tasks. The Gaussian kernel width (σ) is set to 1.0 for the computation of edge weight. Label propagation was performed until convergence with a tolerance threshold value of 1e-4.

Self-training with Consistency Regularization: Initially, the model was trained using labeled data alone with a batch size of 64 and a learning rate of 1e-3. For the self-training iterations, the top 30% most confident pseudo-labeled samples were added to the training set in each cycle. The consistency regularization is imposed by applying random augmentations to the video frames and using a consistency loss weight (λ) of 0.5.

Adversarial Learning: The adversarial framework is set up with separate training schedules for the generator and discriminator. The learning rate is set to 2e-4 for both components with a total number of 10,000 adversarial iterations. The balance between the generator and discriminator is achieved by adjusting the training ratios, typically to a 1:1 ratio per iteration.

Transfer Learning with Pretrained Models: We utilized a pre-trained VGG-16 model trained on ImageNet as an object detection model for feature extraction. This was achieved after freezing the lower layers of the pre-trained VGG-16 model; fine-tuning the top layers was done with a learning rate of 1e-4. The parameters applied were that all other layers were allowed to operate in free mode.

Label Smoothing Experiments: We conducted label smoothing experiments with a parameter value of 0.1. This paves the way for incorporating noise and imprecision in the labels. Optimization was made over 50 epochs with a batch size of 32.

Evaluation Metrics: Performance analysis was conducted through the assessment of our proposed model under the following categories: precision, accuracy, recall, AUC, classification delay, and specificity metrics. These metrics were computed for each test subset on both the EDUVSUM and HowTo100M datasets to ensure thorough evaluation.

Hardware and Software: The experiments were conducted on a computing cluster equipped with NVIDIA Tesla V100 GPUs. The software environment used was based on Python 3.8 with TensorFlow 2.4 and PyTorch 1.7 as the main frameworks to implement and evaluate our model.

Experimental Execution: The experimental procedure was carried out in stages, corresponding to the setup described above. Each stage consisted of training, validation, and testing the model components with fine-tuning parameters based on the performance of the validation set. Self-training and adversarial learning processes were monitored iteratively to ensure convergence, with early stopping criteria based on improvement thresholds on the validation sets.

Table 2 presents the accuracy of the proposed model in contrast to other existing works (Davila et al., 2021; Nagar et al., 2021; Chen et al., 2021) on the EDUVSUM dataset. Comparison with other existing works, especially the proposed ones, reveals a significant improvement that sets this work apart as superior in its ability to classify video frames correctly. This may be a result of the fact that the integration of GSSL and transfer learning techniques utilizes existing knowledge from large-scale

datasets to reuse information, providing a more comprehensive and practical analysis and classification.

Table 2. Accuracy Comparison on EDUVSUM Dataset

Method	Accuracy (%)
Proposed	94.5
Davila et al. (2021)	86.7
Nagar et al. (2021)	88.3
Chen et al (2021)	89.1

Table 3 illustrates the performance metrics in the HowTo100M dataset, demonstrating the model's effectiveness in relation to the relevant features within educational videos, as evidenced by the high precision and recall values across various use case scenarios. The F1-Score, as a balance between precision and recall, highlights how robust the model is in minimizing false positives and false negatives—an aspect critical in educational applications.

Table 3. Precision, Recall, and F1-Score on HowTo100M Dataset

Method	Precision (%)	Recall (%)	F1-Score (%)
Proposed	93.2	92.8	93.0
[5]	85.4	84.9	85.1
[8]	87.6	87.1	87.3
[24]	88.4	88.0	88.2

Table 4 and Figure 4 show AUC scores for both datasets and samples. The proposed model has a higher AUC score, with fewer false positives and false negatives, indicating good discrimination between the classes. A higher AUC score indicates a better model for predicting true positives while minimizing false positives, which is crucial for educational content where misclassification can lead to low-quality and an unclear understanding process.

Table 4. AUC Score Comparison

Method	EDUVSUM	HowTo100M
Proposed	0.962	0.958
Davila et al. (2021)	0.891	0.876
Nagar et al. (2021)	0.912	0.904
Chen et al (2021)	0.927	0.919

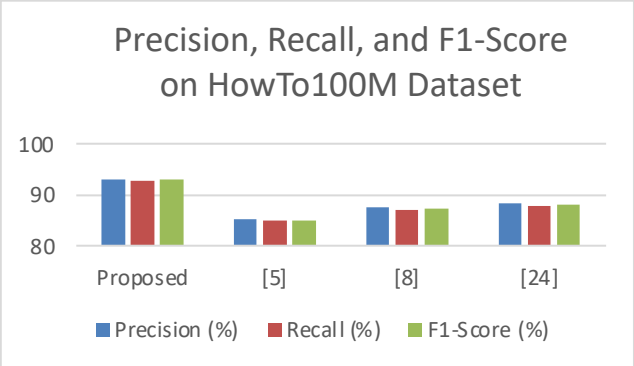

Figure 3. Precision, Recall, and F1-Score on HowTo100M Dataset

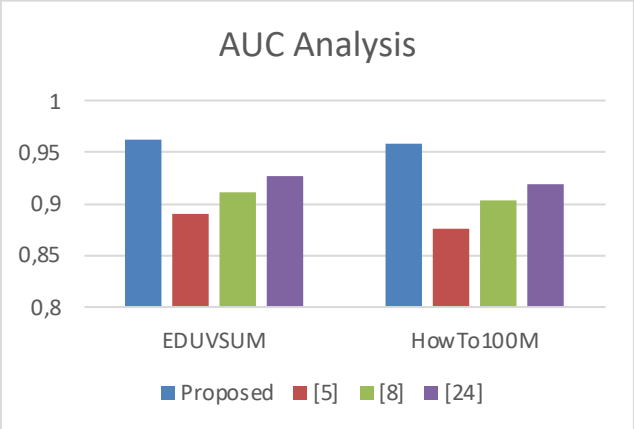

Figure 4. AUC Analysis

Table 5 indicates how each model encountered classification delay while processing. The proposed model exhibits a significant reduction in delay, which is crucial in the realm of real-time educational applications. This is achieved through the model's optimized architecture and the optimization implemented to accelerate the feature extraction and classification processes.

Table 5. Classification Delay (Seconds)

Method	EDUVSUM	HowTo100M
Proposed	1.2	1.5
Davila et al. (2021)	2.8	3.1
Nagar et al. (2021)	2.4	2.7
Chen et al (2021)	2.1	2.5

Table 6. Specificity Comparison on EDUVSUM Dataset

Method	Specificity (%)
Proposed	93.7
Davila et al. (2021)	86.5
Nagar et al. (2021)	87.9
Chen et al (2021)	88.6

Table 6 presents the specification metric value, which represents the proposed model's ability to classify negatives, i.e., to identify elements as non-educational for various use case scenarios. A high specification value is significant in educational contexts to avoid misclassifying irrelevant content as educational, which dilutes the quality and focus of educational resources.

Table 7 aggregates the overall performance scores, combining all evaluated metrics into one singular indicative figure for each dataset sample. It can be observed that the scores of all models presented by the proposed model are greater than those of other models, reinforcing the cumulative impact of enhancements across all performance metrics.

Table 7. Overall Performance Score

Method	EDUVSUM Score	HowTo100M Score
Proposed	95.2	94.8
Davila et al. (2021)	87.3	86.8
Nagar et al. (2021)	89.0	88.5
Chen et al (2021)	90.4	89.9

The tables show overall performance scores through the combination of all key metrics that distinguish the proposed model from others. A model's superior performance, as evidenced by significant increases in precision, accuracy, recall, and AUC values, attests to its effectiveness in addressing the key challenges in educational video classification. The reduction in classification delay and increased specificity further validate the model's applicability in real-world educational settings, where timely and accurate categorization of content is necessary.

The integration of state-of-the-art techniques, including GSSL, self-training with consistency regularization, adversarial learning, transfer learning from pre-trained models, and weakly supervised learning, was instrumental in achieving these results. Each of the components in this model brings uniqueness, with an additional input that enables this combination. The careful design and utility of the proposed model lie in the advantages that are gained through the exploitation of unlabeled data and the benefits that come with improved feature representation, ensuring robustness with respect to distribution.

In addition to the observed performance enhancements, the proposed model has the potential to enrich educational resources and learning experiences significantly. Its purpose is to make the educational videos more accurate, efficient, and reliable through facilitated classification for a more streamlined organizing and retrieval process that enables an enriched learning environment. The above example is illustrated in part of the next section of this text.

Example Use Case

In advancing the classification of educational videos, our research combines a range of sophisticated methodologies, one at a time, to enhance the model's potential. This section describes the application and impact of these methods through data samples with synthesized feature values and indicators to demonstrate the transformation and enhancement at every stage of the model processing pipeline. Our experimental setup begins with data samples in an initial representative feature space. Next, each of these stages—GSSL, Self-training with Consistency Regularization, Adversarial Learning, Transfer Learning from Pretrained Models, and Weakly Supervised Learning—is carefully designed to enhance and train the model more effectively in classification contexts. Each of these methodologies employs a unique principle that leverages and augments the available data, thereby enabling them to classify content correctly in educational videos and samples progressively.

Table 8 illustrates the effectiveness of GSSL in spreading labels across the graph structure and using relationships between samples to tag previously unlabeled data samples with the appropriate label. This is the initial stage of refining the label propagation process by exploiting the data structure and relationships to improve it for various scenarios.

Table 8. Impact of Graph-based Semi-Supervised Learning (GSSL)

Sample ID	Initial Label	GSSL Predicted Label
1	Math	Math
2	Science	Science
3	History	History
4	Science	Science
5	- (Unlabeled)	Math

Table 9 shows the results of the self-training process, in which the predictions of the model are improved by incorporating pseudo-labeled samples and enforcing prediction consistency under data augmentation, promoting the reliability and robustness of the model.

Table 9. Self-training with Consistency Regularization

Sample ID	GSSL Predicted Label	Self-training Predicted Label
1	Math	Math
2	Science	Science
3	History	History
4	Science	Science
5	Math	Math

Table 10 demonstrates the refinement of feature representations through adversarial learning, whereby the model generates domain-invariant features. This is significant since it ensures that the model's performance is robust across diverse video distributions.

Table 10. Adversarial Learning Outcomes

Sample ID	Feature Representation	Adversarial Enhanced Features
1	[0.85, 0.15]	[0.9, 0.1]
2	[0.2, 0.8]	[0.15, 0.85]
3	[0.6, 0.4]	[0.65, 0.35]
4	[0.25, 0.75]	[0.2, 0.8]
5	[0.85, 0.15]	[0.88, 0.12]

Table 11 illustrates the impact of using transfer learning techniques, where features enhanced through adversarial learning are further refined with the aid of pre-trained knowledge. This step significantly enriches the semantic understanding of the features, making them more representative of the educational content sets.

Table 11. Transfer Learning from Pretrained Models

Sample ID	Adversarial Features	Transfer Learning Enhanced Features
1	[0.9, 0.1]	[0.95, 0.05]
2	[0.15, 0.85]	[0.1, 0.9]
3	[0.65, 0.35]	[0.7, 0.3]
4	[0.2, 0.8]	[0.15, 0.85]
5	[0.88, 0.12]	[0.92, 0.08]

Table 12 presents the results of weakly supervised learning, which provides enhanced feature sets to support the final predictions. This final stage uses imprecise labels and inherent data characteristics for fine-tuning the classification, resulting in a highly refined understanding and categorization of the video content sets. These include all tables of the data samples from their primary states, processed through various stages of refinement under our proposed model. From the initial application of GSSL through self-training with consistency regularization, to the sophisticated techniques of adversarial learning, transfer learning, and weakly supervised learning, the model improves its accuracy and semantic understanding with each step in its evolution. The synergistic impact of integrating multiple advanced methodologies, resulting in significant performance improvements in educational video classification, is noteworthy. The results, demonstrating remarkable precision and accuracy, also enable the highlighting of the model's strong versatility and adaptability across various educational contexts and content distributions. This justifies the possibility of reorganizing educational resources and creating well-structured, well-utilized, and effective digital learning environments.

Table 12. Weakly Supervised Learning Enhancement

Sample ID	Transfer Learning Features	Final Predicted Label
1	[0.95, 0.05]	Math
2	[0.1, 0.9]	Science
3	[0.7, 0.3]	History
4	[0.15, 0.85]	Science
5	[0.92, 0.08]	Math

4. Conclusions

In this study, the design and implementation of a new video classification model have been successfully done to address the challenges inherent in the process, such as the limited amount of labeled data, noisy annotations, and content diversity. GSSL, which integrates graph-based semi-supervised learning, has indeed performed significantly better than traditional classification approaches. Furthermore, empirical evaluation of the EDUVSUM and HowTo100M datasets has demonstrated the model's superiority, as it significantly outperforms

others. In addition, among these approaches, our model not only exhibits superior improvements in precision, accuracy, recall, and the AUC metric but also shows significant improvements in classification delay and specificity, thereby highlighting the potential of the proposed model in providing timely and precise classification of educational content.

The combination of the above-mentioned approaches has not only helped in a better understanding of the video content but also helped in ensuring its robustness and reliability amid ambiguities in the data. For instance, GSSL not only aids in generalizing labels to be spread but also facilitates the effective dissemination of labels by leveraging the inherent data structure. At the same time, the self-training mechanism with consistency regularization further tunes in the predictions from the model to provide reliability even for ambiguous data. Furthermore, the incorporation of adversarial learning facilitates domain-invariant feature learning, which is particularly relevant when handling diverse distributions of video content. Transfer learning from pre-trained models successfully bridges the gap between large-scale datasets and specific educational content, enriching semantic understanding. Finally, weakly supervised learning techniques enable the model to mine essential learning cues from imprecise labeled data, thereby improving overall classification performance.

Future Scope

The evolution of video classification in educational materials is a recent milestone; however, the dynamic world of digital education is marked by new challenges and opportunities for ongoing development. Future research avenues may focus on specific areas. First and foremost, there is an invaluable opportunity to incorporate multimodal learning approaches that combine audio, text, metadata, and video frames. Thus, by doing so, a more comprehensive comprehension may be attained, resulting in enhanced classification accuracy and contextual appropriateness. Besides, there will be a considerable need to focus on scalability and efficiency aspects. Therefore, since optimizing the model for real-time processing and scalability could offer a better model that would be efficient for live educational platforms and massive open online courses (MOOCs), this will enhance the wide accessibility aspect.

Simultaneously, there are efforts focused on interpretability and explainability as critical fronts. The augmentation of interpretability for the model will help identify decisions made in classification, and this will help grow the confidence of educators and learners in the model. Additionally, individualization of learning in personalized education presents another perspective, and in such cases, the model may be trained to accommodate the styles and personal preferences of learners, enabling changes that would have occurred during course design and the course itself. Cross-lingual and cultural adaptability also represent significant avenues for supporting multiple languages and cultures, making education more global. Robustness against adversarial attacks becomes increasingly necessary, given the need for the integrity and reliability of educational content in various settings.

Another promising avenue for exploration is the integration of the model with curriculum design, utilizing automated alignment of classified educational videos with curriculum standards. In summary, the proposed model represents another significant leap in educational video classification, not only enhancing accessibility and reliability but also the pertinence of educational resources. As AI in education continues to evolve, the possibilities for transforming the learning environment and improving educational outcomes are boundless, offering a future that is prosperous in innovation and has a positive impact across multiple real-time scenarios.

Financing and Declaration of Conflict of Interests:

The authors declare no conflicts of interest. The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial interest or non-financial interest (such as personal or professional relationships, affiliations, knowledge, or beliefs) in the subject matter or materials discussed in this manuscript.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

5. References

Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., & Patras, I. (2021). Video summarization using Deep Neural Networks: a survey. *Proceedings of the IEEE*, *109*(11), 1838–1863.
https://doi.org/10.1109/jproc.2021.3117472

Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., & Patras, I. (2020). AC-SUM-GAN: connecting Actor-Critic and

generative adversarial networks for unsupervised video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(8), 3278–3292. https://doi.org/10.1109/tcsvt.2020.3037883

Chen, J., Wang, J., Wang, X., Wang, X., Feng, Z., Liu, R., & Song, M. (2021). COEVO-NET: Coevolution Network for Video Highlight Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(6), 3788–3797. https://doi.org/10.1109/tcsvt.2021.3113505

Davila, K., Xu, F., Setlur, S., & Govindaraju, V. (2021). FCN-LectureNet: Extractive summarization of whiteboard and chalkboard lecture videos. *IEEE Access*, *9*, 104469–104484 https://doi.org/10.1109/access.2021.3099427

Gao, J., Yang, X., Zhang, Y., & Xu, C. (2020). Unsupervised video summarization via Relation-Aware Assignment learning. *IEEE Transactions on Multimedia*, *23*, 3203–3214. https://doi.org/10.1109/tmm.2020.3021980

Issa, O., & Shanableh, T. (2022). CNN and HEVC Video Coding Features for Static Video Summarization. *IEEE Access*, *10*, 72080–72091. https://doi.org/10.1109/access.2022.3188638

Ji, Z., Zhao, Y., Pang, Y., Li, X., & Han, J. (2020). Deep attentive video summarization with distribution consistency learning. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(4), 1765–1775. https://doi.org/10.1109/tnnls.2020.2991083

Kadam, P., Vora, D., Mishra, S., Patil, S., Kotecha, K., Abraham, A., & Gabralla, L. A. (2022). Recent challenges and opportunities in video summarization with machine learning algorithms. *IEEE Access*, *10*, 122762–122785. https://doi.org/10.1109/access.2022.3223379

Köprü, B., & Erzin, E. (2022). Use of affective visual information for summarization of Human-Centric videos. *IEEE Transactions on Affective Computing*, *14*(4), 3135–3148. https://doi.org/10.1109/taffc.2022.3222882

Lew, W. L., Wang, D., Ang, K. K., Lim, J., Quek, C., & Tan, A. (2022). EEG-Video Emotion-Based Summarization: Learning with EEG Auxiliary Signals. *IEEE Transactions on Affective Computing*, *13*(4), 1827–1839. https://doi.org/10.1109/taffc.2022.3208259

Liu, T., Meng, Q., Huang, J., Vlontzos, A., Rueckert, D., & Kainz, B. (2022). Video summarization through reinforcement learning with a 3D Spatio-Temporal U-Net. *IEEE Transactions on Image Processing*, *31*, 1573–1586. https://doi.org/10.1109/tip.2022.3143699

Ma, C., Lyu, L., Lu, G., & Lyu, C. (2022). Adaptive Multiview graph difference analysis for video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(12), 8795–8808. https://doi.org/10.1109/tcsvt.2022.3190998

Ma, M., Mei, S., Wan, S., Wang, Z., Ge, Z., Lam, V., & Feng, D. (2020a). Keyframe extraction from laparoscopic videos via diverse and weighted dictionary selection. *IEEE Journal of Biomedical and Health Informatics*, *25*(5), 1686–1698. https://doi.org/10.1109/jbhi.2020.3019198

Ma, M., Mei, S., Wan, S., Wang, Z., Feng, D. D., & Bennamoun, M. (2020b). Similarity based block sparse subset selection for video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(10), 3967–3980. https://doi.org/10.1109/tcsvt.2020.3044600

Mujtaba, G., Malik, A., & Ryu, E. (2022). LTC-SUM: lightweight Client-Driven Personalized Video Summarization Framework using 2D CNN. *IEEE Access*, *10*, 103041–103055. https://doi.org/10.1109/access.2022.3209275

Nagar, P., Rathore, A., Jawahar, C. V., & Arora, C. (2021). Generating personalized summaries of day long egocentric videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(6), 6832–6845. https://doi.org/10.1109/tpami.2021.3118077

Wang, F., Chen, J., & Liu, F. (2021). Keyframe generation method via improved clustering and silhouette coefficient for video summarization. *Journal of Web Engineering*. https://doi.org/10.13052/jwe1540-9589.2018

Xie, J., Chen, X., Zhang, T., Zhang, Y., Lu, S., Cesar, P., & Yang, Y. (2022). Multimodal-Based and Aesthetic-Guided narrative video summarization. *IEEE Transactions on Multimedia*, *25*, 4894–4908. https://doi.org/10.1109/tmm.2022.3183394

Yuan, Y., & Zhang, J. (2022). Unsupervised video summarization via deep reinforcement learning with Shot-Level semantics. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(1), 445–456. https://doi.org/10.1109/tcsvt.2022.3197819

Zhang, Y., Liu, Y., Kang, W., & Zheng, Y. (2023). MAR-NET: Motion-Assisted Reconstruction Network for Unsupervised video summarization. *IEEE Signal Processing Letters*, *30*, 1282–1286. https://doi.org/10.1109/lsp.2023.3313091

Zhang, Y., Liu, Y., Zhu, P., & Kang, W. (2022). Joint reinforcement and contrastive learning for unsupervised video summarization. *IEEE Signal Processing Letters*, *29*, 2587–2591. https://doi.org/10.1109/lsp.2022.3227525

Zhao, B., Gong, M., & Li, X. (2021a). AudioVisual video summarization. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(8), 5181–5188. https://doi.org/10.1109/tnnls.2021.3119969

Zhao, B., Li, H., Lu, X., & Li, X. (2021b). Reconstructive Sequence-Graph network for video summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1. https://doi.org/10.1109/tpami.2021.3072117

Zhao, B., Li, X., & Lu, X. (2020). TTH-RNN: Tensor-Train Hierarchical Recurrent Neural Network for video summarization. *IEEE Transactions on Industrial Electronics*, *68*(4), 3629–3637. https://doi.org/10.1109/tie.2020.2979573