I. J. Engineering and Manufacturing, **2023, 3, 10-24**
Published Online on June 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijem.2023.03.02

A Comparative Analysis of Video Summarization Techniques

Darshankumar D.Billur*
KLE Collegeof Engineering & Technology/ Department of ECE, Chikodi-591201, India
Email: darshankumar999@gmail.com
ORCID iD: https://orcid.org/0000-0001-5765-947X
*Corresponding Author

Manu T. M.
KLE Institute of Technology, Hubballi / Department of ECE, 580030, India
Email: manutmece@gmail.com
ORCID iD: https://orcid.org/0000-0002-6091-1098

Vishwas Patil
KLE Institute of Technology, Hubballi / Department of ECE, 580030, India
Email: v.bkl56@gmail.com
ORCID iD: https://orcid.org/0000-0002-4127-1726

Received: 08 September, 2022; Revised: 20 October, 2022; Accepted: 16 November, 2022; Published: 08 June, 2023

Abstract: Video summarization special field of signal processing which includes pre-processing of video sets, their contextual segmentation, application-specific feature extraction & selection, and identification of dissimilar frame sets. Various variety of machine learning models are proposed by researchers to design such summarization methods, and each of them varies in terms of their functional nuances, application-specific advantages, deployment specific limitations, and contextual future scopes. Moreover, these models also vary in terms of quantitative & qualitative measures including accuracy of summarization, computational complexity, delay needed for summarization, precision during the summarization process, etc. Due to such a wide variation in performance levels, it is difficult for researchers to identify optimal models for their functional-specific &performance-specific use cases. Because of this, researchers and summarization-system-designers are required to validate individual models, which increases the delay & cost needed for final model deployments. To overcome these delays & reduce deployment costs, this paper initially discusses a multiple variety of video summarization models in terms of their working characteristics. Based on this discussion, researchers shall be able to identify optimum models for their functionality-specific use cases. This paper also analyzes and compares the reviewed models in terms of their performance metrics including summarization accuracy, delay, complexity, scalability and fMeasure, which will further allow readers to identify performance-specific models for their deployments. A novel Summarization Rank Metric (SRM) is calculated based on these evaluation metrics, which will assist readers to identify models that can perform optimally w.r.t. multiple evaluation parameters & different use cases. This metric is calculated by combining all the comparison metrics, which will assist in identification of models that have high accuracy, low delay, low complexity, high scalability & fMeasure levels.

Index Terms: Video, Summarization, Deep Learning, Bioinspired, SRM, Accuracy, Delay, Complexity, Scalability

1. Introduction

In current age of digital technology, video is a media that allows for the recording, reproduction, replaying, transmission, and display of moving visual material. People are more likely to recall information when it is presented visually as opposed to verbally; this is one of the reasons why videos outperform all other forms of media in terms of viewership. Due to the extensive availability of the internet in the modern day, every one of us may readily locate a wide variety of movies that meet their own likes and needs. Researchers have a limited amount of time, which makes it tough for us to see the lengthy films that researchers deemed valuable despite their length. Video summaries make it possible to construct an educational video in less time than would be possible without them. A video with a short

runtime that contains all of the content from the original video which is much moreefficient and cost-effective for the end user. This is true when both time and money are considered. Researchers are able to do video indexing using the summarized videos, hence accelerating search operations in the multimedia database. In this field of research, several methodological techniques have been used. Some researchers created summary algorithms based on aural characteristics, score boxes, and goal identification. With the use of restricted satisfaction programming, while others created systems through which a user may describe films based on his or her personal preferences. Users may add inputs or constraints into summarizing programs, and the software constructs frames based on these inputs. These kinds of programs have led to the creation of software for summarization. Some approaches seem to be more effective than any other clustering technique when it comes to deleting crucial video frames. CSSP, also known as the column subset selection problem, is used to pick the most appropriate keyframes from the population among set of relevant frames that CSSP gives for summarization.

The subfield of signal processing known as video summarization involves the reprocessing of video sets, the contextual segmentation of those sets, the extraction and selection of application-specific features, and the detection of distinct frame sets. In order to build such summarizing approaches, researchers have presented a wide range of machine learning models, each of which differs in terms of its functional subtleties, application-specific benefits, deployment-specific constraints, and contextual future horizons. In addition, these models differ in terms of quantitative and qualitative variables such as the accuracy of the summarizing, the computing complexity, the latency required for the summarization, the precision throughout the process of summarization, and so on. Researchers have a tough time locating models that are best for the functional-specific and performance-specific use cases they have developed since there is such a large range in the degrees of performance. As a result of this, researchers and the designers of summarization systems are required to evaluate each model, which both increases the amount of time necessary for final model deployments and the associated costs.

In order to get around these limitations, and cut down on the expenses associated with deployment, the first part of this article analyzes a number of different video summarization models and the working features of each of them. Researchers should now be able to determine the best models for the functionality-specific use cases they are investigating as a result of this conversation. This paper also analyzes and compares the models that were reviewed in terms of their performance metrics such as summarization accuracy, delay, complexity, scalability, and fMeasure. This will allow readers to further identify performance-specific models for their deployments by providing them with these information sets.

It can be observed that, deep learning technology is widely used in variety of automated video summarizing applications. Their media assets may now be properly indexed so that they can be searched through and promoted. With this, viewers are able to have a better watching experience while also boosting their consumption of material. Movie trailers and teasers are also generated using video summaries that are tailored to the specific needs of the different content presentation situations. A video summary of the most important moments caught by security cameras may be created using this technique to emphasize the most important moments of an event. This technique might be useful in both of these scenarios. This is a win-win situation for both time-saving development monitoring and security. As a result of these benefits, researchers and network engineers are increasingly turning to video summarizing models based on deep learning. This trend might be attributable to the above-mentioned advantages. Readers will be given a brief overview of these methods in the next section. Following that, the efficacy of these regimens will be extensively studied in contrast to those previously mentioned. The result is that video summary system researchers and developers will be able to choose the optimum mix of protocols to provide the highest level of privacy protection for each of their specific deployments, as a result. The paper finishes with some thoughtful observations and suggestions on how to enhance the models that have been examined which assists in improving their real-time performance levels. The comparison of the models' performance metrics, such as summarization accuracy, delay, complexity, scalability, and fMeasure, in this paper will help readers choose models that are performance-specific for their deployments. These evaluation metrics are used to calculate a novel Summarization Rank Metric (SRM), which helps readers find models that perform well given a variety of evaluation criteria and use cases.

2. Literature Review

A wide variety of Video Summarization models are proposed by researchers, and each of them vary in terms of the internal working configurations, and their real-time performance characteristics. In this section, a theoretical survey of these models in terms of their functional nuances, application-specific advantages, deployment-specific limitations, and contextual future scopes is described, which will assist readers to identify functionality specific summarization models for their deployments. Work in [1] observed that video summarization (VSUMM) is a widely used technique for handling enormous volumes of video data. To appropriately represent the subject matter of the video, VSUMM chooses pertinent still images. Currently, approaches can only extract still images from videos as content summaries, excluding any information about motion. To address these issues, an efficient and unique framework for summarizing video data and motion is presented. Using Capsules Net, which has been trained to extract spatial and temporal data, a motion curve between frames is produced. Next, researchers demonstrate how to automatically trim movies into shots using transition effects. Researchers are able to identify key-frame sequences inside certain frames using a novel way of Self-

Attention (SA), choose still images to summarize video content, and maybe estimate optical flows to offer a comparable service for video motions. The final experimental findings on the VSUMM, TvSum, SumMe, and RAI datasets demonstrate how competitive their technique is for shot segmentation and video content summary.Work in [2] asserts that sophisticated video summarizing algorithms may exclude redundant material while maintaining the most relevant and instructive portions of the original movie. Researchers present the 3DST-UNet-RL video-summarization system. For future RL, input video data is encoded using a 3D spatio-temporal U-Net (RL). An RL agent must first study the spatio-temporal latent scores in order to predict whether a video frame will be saved or destroyed. It's likely that real/inflated 3D spatiotemporal CNN features perform better when learning video representations than 2D image features. They assert that both supervised and unsupervised scenarios could be compatible with their method. The authors analyze the effects of proposed summary lengths while demonstrating 3DST-UNet-efficacy RLs on two benchmarks. The group used their approach to the assignment of summarizing surgical video. When reviewing patient video data as part of retrospective analysis or audit, the suggested video summarizing approach has the potential to lower ultrasonic screening movie storage costs and boost efficiency.

Large movie dataset management, storage, and indexing issues [3] need to be addressed immediately. While most videos lack text descriptions required for semantic analysis approaches, low-level characteristic models are unable to synthesize essential semantic information. This makes video semantic mining more practical. The use of action parsing and reinforcement learning to summarize videos is suggested. The model uses a combination of RL-based video summarization with action-based video segmentation and parsing. Using annotated data, researchers first train a sequential multiple instances learning model to address the long and hazy comprehensive annotation issue. After that, they use deep recurrent neural networks to create a video summary model. Their classification accuracy reveals the effectiveness of the key frame identification and classification process. Experimental results and comparisons to other approaches have shown the benefits of the suggested method.In [4], a novel method for locating still video keyframes was created. During the first processing of the dataset, researchers employed HEVC encoding. 64 characteristics were suggested for each frame during coding. The YUVs were transformed to RGBs by the researchers before submitting them to CNN networks that had already been trained to extract features from raw videos. Comparing VGG16, Inception-ResNet-v2, AlexNet, GoogleNet, and the latter. Researchers may now choose to utilize updated data sets. Before trying to identify keyframes, it is crucial to remove similar video frames. These stills were identified and removed from the movie using a limited number of HEVC's characteristics. The researchers provide an elimination procedure that considers the absolute differences between two motion-compensated frames. Comparisons are made between the suggested methods and current SIFT flow algorithms that use CNN characteristics. Before the critical frames were identified, stepwise regression was employed to potentially reduce the dimensionality of the feature vectors. Existing works achieve dimensionality reduction utilizing sparse autoencoders using CNN features.

The bulk of network traffic now comes from mobile edge device videos because to advancements in wireless multimedia. A video summary is provided [5] to help viewers decide whether or not to watch a movie entirely before getting started. Due to the limits of edge devices and dynamic wireless networks, a user-oriented and adaptive solution is required for video summarizing and distribution. Researchers should produce and deliver a video summary depending on network connections and the user's tolerance for delay in order to optimize the user experience and bandwidth utilization ratio. Previous summarizing techniques lacked the adaptability to change the summary size according on the bandwidth available or the user's tolerance for delay. In order to enable quick and flexible processing of mobile videos, the video summarization optimization task is formulated while accounting for the elasticity of the number of representative segments and outlier identification. Elastic Video Summarization Algorithm, an online greedy algorithm, was developed by academics to handle NP-hard problems (EVS). Researchers have investigated its features and created EVS-II in an effort to lessen the computational complexity of EVS. The results of the studies confirm the effectiveness of their suggested methods for matching network bandwidth and detecting anomalies.Less invasive chest procedures used for lung cancer and asthma are now feasible thanks to bronchoscopy. A clinician may inspect the intricate 3-D airway tree, collect tissue samples, and diagnose or treat a patient's ailment using the bronchoscope's video feed. As a result of information overload, doctors often discard useful procedural videos. Analogies and memories are used to chart the development of a technique. The possibility of automatically summarizing endobronchial video streams is suggested. The approach in [6] leverages biomedical video sets and includes the three elements of shot segmentation, motion analysis, and keyframe selections (SMK). This technique can provide a shot list and keyframes for a procedural video. Contrary to earlier techniques, endoscopic video recording offers a uniform description of spontaneous, unedited material. Their approach effectively covers endobronchial areas and is resistant to changes in parameter values. Their method only utilised 6.5% of the available video frames to obtain 92.7% video coverage across a variety of video sequences. A system that is presently being developed uses the generated video summary to enable direct connection with a patient's 3-D chest CT scan and to enable quick video browsing and retrieval across intricate airway tree models.

Summarizing videos helps you locate what you're looking for while saving time and money [7]. The majority of the video summarizing algorithms that are now on the market do not reliably explain the events shown in the video, including the nonchronological movement of objects or the size and placement of backdrops and foregrounds. In this study, researchers provide a method for creating a trustworthy video summary. To identify crucial video actions, their solution first uses a static and dynamic scene optimization framework (SDSOF). Second, the obvious activity frames are fixed using alpha matting (AM). Third, the information is obtained quicker and requires less storage space since the video frames are condensed into a single frame. The study provided here suggests that a single-frame indexing strategy

is better than a multi-frame indexing approach for the video database. In a series of tests, the suggested approach is put to the test. According to the findings of the studies, their solution is effective in terms of calculation time and memory utilization while still producing a strong visual summary.With the use of video summarizing, researchers may more easily watch, catalog, and retrieve a large number of videos [8]. The objective of video summary is to accurately convey the plot of the original using the fewest possible frames. Unsupervised video summarizing techniques are easier to use since they don't need a lot of key frame annotation to train a model. The fact that some information is lost in the conversion from the original to the summary reduces the usefulness of videos. To minimize data loss during video summarization, the authors of this study created a unique Cycle-consistent Adversarial LSTM (CCA LSTM) architecture. Cycle-evaluations are built on the principles of frame-consistent learning. SUM's A bidirectional LSTM network detector is used to find the connections between video frames. A standard for how much of the original video should be included in the summary version may be established by the evaluator using the "supervise" option. One GAN learns to rebuild the original video from the summary video, while the second GAN learns to do the reverse, and the two GANs work together to do the assessment. Researchers have connected cycle learning with the maximization of mutually useful information in order to standardize summarization. Extensive tests on three benchmark datasets for video summaries show that Cycle-superiority SUM outperforms other unsupervised approaches.

Work in [9] observed that a rise in the use of video lectures that have been recorded and distributed is substantial due to boost in online learning scenarios. Researchers have discovered recordings of math lectures and seminars covering a broad variety of levels. Many of these concepts may be best shown on chalkboards and whiteboards. There are many recordings of lectures that all include the speaker taking notes. The research team suggests a unique technique for extracting and summarizing handwritten video material. When presented with a video that contains handwriting, FCN-LectureNet can extract the text as binary images. They are put to the test to determine unique and trustworthy units of handwritten information in order to establish a spatial-temporal index. The spatial-temporal index is used to provide a signal that may be used to roughly estimate the removal of material. In order to construct chronological lecture segments based on the assumption that lecture subjects change as extraneous material is weeded out, the highest points of this signal are utilised. Researchers utilize the segments to manually write a key-frame summary of the data. As a consequence, the accessibility of lecture video material will be improved. To construct Lecture Math for this study, researchers enhanced the Access Math datasets. In experimental findings on both datasets, but notably on the bigger and more complex dataset, their unique technique outperforms accepted standards.Work in [10] addresses supervised video summarizing as a sequence-to-sequence learning task with an input stream of raw frames of video and an output stream of key frames. To mimic the way people-choose significant shots in a video, their technique entails creating a deep summarization network. In a new architecture for video summarizing known as attentive encoder-decoder networks for the summarization of video, the encoder employs a bidirectional long short-term memory (BiLSTM) to retain contextual information among the incoming video frames (AVS). The decoder uses multiple stages of attention-based LSTM networks inclusive with additive and multiplicative goal function respectively. The industry standards for video summarizing datasets are SumMe and TVSum. The results show that AVS-based approaches outperform state-of-the-art techniques for both datasets.

Study in [11] discovered Automatic video summary is on the rise and focuses on the most interesting and pertinent segments of a video. The subjectivity involved makes the task more difficult. Previous studies that relied on costly human annotations or manually established criteria often fell short of expectations. According to experts, a video's related metadata—including the titles, questions, descriptions, comments, and so on—displays human-curated semantics. Current techniques frequently exclude critical background information when describing videos. This study creates a DSSE model to analyze contextual data and summarize videos. The DSSE is used to correlate the hidden layers of two uni-modal autoencoders, each of which uses video frames and other data. By interactively limiting losses of semantic relevance and feature reconstruction, it may be possible to discover common information across video frames and auxiliary information. Results of attempts to measure meaning have been inconsistent. In the end, semantically significant video segments are chosen by decreasing their distances from latent side information. On both the Thumb1K and TVSum50 datasets, tests have shown that DSSE outperforms alternative video summarization techniques.Researchers [12] offer a novel approach based on dynamic network modelling for video summarization (DNMVS). Deep models that were trained on ImageNet are used in the great majority of video summarizing techniques. They were able to record spatial-temporal linkages by using data at the object and relation levels. Their approach produces spatial graphs using object suggestions. From a series of geographic graphs, the researchers create a temporal graph. In order to predict score prediction and selection of key shot, researchers have extracted spatial-temporal representations, as well as relationship reasoning across spatial and temporal graphs, using graph convolution networks. Researchers created a self-attention edge pooling module to lessen relation clutters caused by closely connected nodes. The TVSum and SumMe standards are evaluated by scientists. Testing has shown that the suggested technique performs noticeably better than the top-performing video summarizing methods already in use.

The difficulty in establishing how a video's content connects to the summary being prepared makes video summarizing difficult [13]. The authors of this research provide a three-stage supervised video summarizing procedure that makes use of deep neural networks. The system divides the challenging work of 3D video summarization into manageable computational subtasks using a divide-and-conquer approach. These subtasks are then addressed in a hierarchical manner by use of 2D CNNs, 1D CNNs, and long short-term memory (LSTM). It is more effective and efficient to use a hierarchical spatial-temporal model (HSTM). The authors provide a user-ranking approach to reduce

the inherent subjectivity of labelling in user-generated video summaries and improve labelling quality for strong supervised learning. On two independent benchmark datasets, their strategy outperforms cutting-edge video summarizing techniques.Laparoscopic videos are used in medical teaching and quality control since laparoscopy is so popular for less invasive therapies. Laparoscopic video collections are underused since viewing so many movies would take too much time. Laparoscopic recordings may take a while to manually extract keyframes; however, this research offers a video summary approach that can accomplish so in a matter of seconds. Initially, deep features are produced using convolutional neural networks (CNNs) [14] and used to represent video frames rather than low-level data. Second, based on this deep representation, laparoscopic video summarization is structured as a diversified and weighted dictionary selection approach. To minimize duplication, a diversity regularization term is included, and high-quality keyframes are chosen while taking into consideration image quality. Finally, a quick-convergent iterative approach for model optimization is created and studied. The suggested strategies successfully outperform a newly made available laparoscopic dataset. The suggested approach may improve access to surgery, junior clinician education, patient comprehension, and case file preservation.

The authors of [15] propose a sequence-to-sequence learning paradigm for supervised video summarization, where the sequences of video frames are at input and output with their projected significance ratings. Uneven distribution and a lack of long-term contextualization are investigated. Because they rely on long-term encoder-decoder attention, current approaches have problems obtaining short-term contextual attention information inside a sequence of video. The ground-truth sequence may include inconsistencies that result in an insufficient solution. To address the first issue, researchers added a short-term self-attention mechanism to the encoder. Researchers suggest the use of deep video summarization models as a remedy, which necessitates the collaboration of an encoder and a decoder. The second approach looks for a distribution that works for both sequences by using a simple but effective regularization loss term. Their most effective strategy is a careful video summary that maintains the regularity of dissemination (ADSum). Through rigorous testing on benchmark data sets, ADSum has been shown to be superior than alternative techniques.Work in [16] describes a method for automatically and without human involvement summarizing videos. The proposed architecture uses sequence generation to choose relevant video chunks for the summary and integrates an actor-critic model within a generative adversarial network. The Discriminator then rewards the Actor and the Critic for their efforts after they have gone through many rounds of selecting key-fragments from a video. The two characters, Actor and Critic, could stumble onto a set of possible actions before automatically picking up on a key-fragment selecting technique. The suggested guidelines for selecting the best model after training enable automated selection of optimal values for non-learned training parameters (such as the regularization factor). Experiments on multiple benchmark datasets demonstrate that the proposed AC-SUM-GAN model is able to consistently outperform unsupervised approaches and is competitive with supervised methods (SumMe and TVSum).

The work in [17] generates query-focused video summaries from user inquiries and lengthy videos. Analysis of the intent of user's and the significance of semantic content in the generated summaries is the aim of summarization which focussed on query. Researchers suggest using a QSAN, or a query-biased self-attentive network, to solve this problem. Using semantic (semantic data) information from video descriptions, they want to develop a general summary, which will then be combined with query information to provide a summary that is suited to particular inquiries. Researchers first suggest a hierarchical self-attentive network to express the connection between the various frames within a segment, across segments of the same video, and between the descriptive language and visual elements of the video. To produce informative videos, researchers use a reinforced caption generator that has been tuned using a dataset of video captions. Researchers create a query-aware scoring module to calculate the score for each image depending on the provided query in order to provide the query-focused summary. Numerous tests on the reference dataset show that their strategy is competitive.The growing usage of video summarization (VS) in many applications of computer vision, such as video search, indexing, and browsing, may be the cause of its increased popularity [18]. Via the use of the best features and clusters, VS algorithms have traditionally been improved through research. Due to the growing density of visual sensor networks, processing time and summary quality are trade-offs. Giving a thorough video summary may be difficult since IoT monitoring networks sometimes lack funding. In order to summarize surveillance footage obtained by IoT applications, this study proposes a deep convolutional neural network (DCNN) architecture with hierarchical weighted fusion. They start by extracting CNN components that may be utilized to divide a shot. Researchers use a trained CNN model to predict visual memorability, aesthetics, and entropy in order to make the summary interesting and diverse. Finally, a hierarchical weighted fusion approach is used to aggregate feature scores. An attention curve describing the video is then created using the data. In experimental settings employing benchmark data sets, their technique has been shown to be superior than the state-of-the-art methods.

The study suggested in [19] found every day, a tremendous amount of audio, visual, and textual movies are made. This tendency has been influenced by the growing accessibility of recording applications for smartphones, tablets, and cameras. Understanding visual semantics and converting it into a compacted format, such as a heading or title or summary, is crucial for saving space and assisting users' capacity to index, explore, and absorb information more rapidly. Google created summaries of videos in both plain English and abstract text using their Abstractive Summarization of Video Sequences (ASVS) approach. The viewer is given a written video explanation and abstract summary in this manner, allowing them to choose the most important elements. According to their study, the combined model can outperform standard methods for jobs requiring in-depth, succinct, and understandable multi-line video explanations and summaries.While the research conducted in [20] revealed Summarizing, analysing, indexing, and

retrieving the enormous amounts of video data produced by industrial surveillance networks presents challenges. Multiview video summarizing (MVS) is difficult because to data amount, redundancy, overlapping views, different lighting conditions, and interview correlations. MVS cannot be completely used by soft computing methods that depend on clustering or other basic properties. Researchers provide a two-tiered deep neural network soft computing method for accomplishing MVS in this research. Before transmitting the photos to the cloud for further analysis, the initial online layer divides them apart based on how the target appears. The lookup table's frames' deep characteristics are then sent into Layer 2's DB-LSTM in order to boost informativeness probability and provide a summary. Their solution surpasses cutting-edge MVS techniques, according to an experimental assessment utilizing the benchmark dataset and YouTube industrial surveillance datasets.

In [21], a DaS architecture for dense video captioning is suggested. The researchers then use the current image/video captioning technique to extract visual features (such as C3D features) from each segment and describe it in a single sentence. The researchers first divide each uncut long video into multiple proposals of an event, each of which includes a collection of brief video segments. On the grounds of generated sentences contain rich semantic descriptions of the entire proposal of an event, this study suggests a novel two-stage Long Short-Term Memory (LSTM) approach with a novel hierarchical attention mechanism to summarize all generated sentences as a single descriptive sentence using visual features. By using all of the semantic words from the produced sentences and all of the visual traits from all of the event proposal segments as inputs, the first-stage LSTM network serves as an encoder to efficiently summarize semantic and visual information. The second-stage LSTM network decodes the output of the first-stage network as well as every video segment in an event proposal to produce a descriptive phrase. Experiments using ActivityNet Captions show how useful the DaS architecture is for high-density video captioning.Teachers and parents may be able to assess their children's participation by using emotional analysis of classroom footage. High-definition cameras may now be used to capture lectures in video format. It takes time to watch movies, and it might be difficult to determine whether or not characters' feelings are common or not. The analytics system Emotion Cues [22] combines visuals and emotion detection algorithms to assess classroom recordings. It has three synchronized perspectives: a character view that depicts a single person's distinct emotional state; a summary perspective that highlights the wide range and dynamic development of emotions; and a video perspective that makes it easier to examine video information in-depth. Face size and occlusion are two variables that may have an impact on the accuracy of emotion identification. They provide suggestions about prospective issues and their causes. The suggested system may be able to decipher the emotions in instructional videos, according on the findings of two use cases and interviews with end users and subject matter experts.

Work in [23] claims that while constructing summaries, current video summarizing algorithms give priority to sequential or structural video data. They ignore the video that outlines accountability. The authors of this paper suggest adopting a meta-learning strategy they call MetaL-TDVS to examine the video summarization process in various videos. By reformulating the task as a meta learning problem and fostering the generalization skills of the trained model, MetaL-TDVS aims to reveal the latent video summarizing process. MetaL-TDVS views video summarization as a distinct activity in order to make greater use of existing knowledge and expertise. MetaL-TDVS makes guarantee that the model is consistently accurate across all training videos, not just the one it was trained on, by using two-stage backpropagation for model updates. Numerous tests on benchmark datasets have shown the TDVS and generalization capabilities of MetaL.Recurrent neural networks (RNNs) have improved significantly at summarizing videos, however there are still certain challenges, according to [24]. This essay focuses on two unsolvable problems with RNN video summarization. Massive feature-to-hidden matrices first and foremost. Since video features often exist in high-dimensional spaces with large feature-to-hidden mapping matrices, training an RNN model may be challenging. The second is a failure to pay attention to the long-term temporal dependencies. The thousands of frames in most videos are too much for traditional RNNs to handle. Researchers created a TTH-RNN for video summarization in response to these two problems. It avoids the requirement for large feature-to-hidden matrices by using a tensor-train embedding layer and a hierarchical RNN to assess the temporal correlations between video frames. The experimental findings on SumMe, TVsum, MED, and VTW demonstrate the TTH-superior RNN's performance in video summarization sets.

According to work in [25], companies require efficient video summarizing (VS) techniques to handle the surge of fresh video material in order to preserve and manage vital video information. Industrial movies are difficult to watch online because of the variety and complexity of the events they depict. Thanks to the efforts of the research community, there is now an internet system that can gather videos intelligently, cut out redundant material, and produce summaries. Researchers first capture video data using constrained devices connected to an industrial Internet of Things network that include vision sensors, and then they filter out redundant information by analysing the data's low-level properties. Then, keyframes (KF) are selected by extracting sequential information from the frames before they are transferred to the cloud for further analysis. Researchers find candidate keyframes to identify the most succinct data representations. This article offers an overview of important data as well as a technique for enhancing video data on devices with limited capabilities. Experimental findings on datasets made accessible to the public show a 0.3-unit boost in F1 score in favor of lowering temporal complexity. Work in [26] presented a DSNet model for guided video summarization. There are several DSNet configurations, both with and without anchors. While anchor-free approaches do not depend on any pre-defined temporal suggestions, anchor-based methods employ anticipated significance ratings and segment placements to assist users in finding and selecting representative material from video sequences. Their interest detection methodology, in contrast to previous supervised video summarizing techniques, is the first to make an effort at temporal consistency using the temporal interest detection concept. The anchor-based technique is used to forecast location regression and

significance once a wide sample of temporal interest ideas across a variety of time scales is initially provided. To ensure that the final summary is accurate and thorough, it is crucial to separate the data into positive and negative sections. The disadvantages of temporal proposal are avoided by the anchor-free technique, which predicts video frame and segment importance directly. The architecture might be a component of supervised video summarizing software that is for sale. Researchers utilize SumMeand TVSum to compare the effectiveness of anchor-based versus anchor-free summarization in order to decide which approach is better. Both anchor-based and anchor-free techniques are supported by experiments.

According to [27], dictionary selection with self-representation and sparse regularization (DSSR2) has shown encouraging results when VS is seen as a sparse selection operation on video frames. Dictionary selection techniques in their current iteration entirely ignore the inherent structured data present in video frames and are only effective for data reconstruction. Additionally, the L2 norm's restriction on sparsity is inadequate, which results in a large number of duplicate keyframes. The authors of this paper provide a generic framework for graph convolutional dictionary selection with L2p norm for keyframe selection and skimming-based summarization (GCDS). Researchers initially include graph embedding into dictionary selection before considering structured video data. The authors of the paper suggest two methods that make use of L2p norm constrained row sparsity to summarize videos. Use p=0 to choose a range of sample keyframes, whereas p=1 is best for rapid scanning purposes. Theoretically, an iterative technique to improving the suggested model converges. The outcomes of tests performed on four reference data sets show that the suggested procedures are better and effective.Video sensors are widely employed for smart transportation and security applications, according to research [28]. However, owing to constrained storage and transmission capacity, real-time Big Data processing is difficult. Unneeded video sensor data is filtered out using MVS. Existing MVS techniques need additional streaming to finish summarization, a large amount of bandwidth, and are incompatible with industrial IoT. This is due to the fact that they analyze video data offline by transferring it to a local or cloud server (IIoT). In this paper, researchers provide a convolutional neural network and Internet of Things-based CI MVS architecture. They use Raspberry Pi (RPi) smart devices (clients and master) with built-in cameras as part of their strategy to record video from various perspectives. Each Raspberry Pi (RPi) client employs a lightweight convolutional neural network (CNN) model to recognize items in photos, evaluate the scene for traffic and population density, and look for problematic things to find alerts for the IIoT networks.

Work in [29] noted that character-focused video summaries are becoming more and more popular as "re-creation" tendencies spread throughout social media platforms. Regular human searches would need a lot of resources, but automated extraction is laborious and misses often. Two types of textual material that often accompany videos on social networking sites are subtitles and bulleted screen remarks. The inclusion of textual material may be beneficial for character-focused video summaries. In this study, researchers provide a novel framework for encoding both textual and graphical data. This is accomplished by using detection techniques to discover characters at random, followed by re-identification to determine which characters to concentrate on in order to extract likely key-frames, from which pertinent textual material is automatically selected and incorporated based on frame attributes (Char KF). A character-focused summary will be created from these critical moments. Through testing on actual data, it is shown that their method outperforms state-of-the-art baselines on character-oriented video summary problem sets.According to research published in [30], comprehensive video captioning has made it possible to identify and caption every occurrence in a lengthy, unedited video. Even while findings are encouraging, most current systems struggle to keep up when scenes and objects change over the course of a lengthy proposal because they don't carefully consider scene evolution within an event temporal captioning proposal. The researchers suggest a two-stage graph-based segmentation and summarizing method to deal with complicated video captioning. For more accurate captioning, video suggestions are "partitioned" into progressively smaller chunks. The process of integrating the various sentences describing each segment into a single statement that gives a broad perspective of the whole event is referred to as "summarization." A new paradigm based on the interconnection of words has been advocated by academics, who have focused on the idea of "summarization." Researchers treat semantic words as nodes in a network and infer their connections to visual inputs to accomplish this objective. For GCN-LSTM Interaction (GLI) modules, one of two techniques is recommended. Their approach succeeds well on YouCook II and ActivityNet Captioned datasets.

Video summarizing (VS) may be seen as a subset selection task where representative keyframes alias key segments are chosen from a whole video frame collection, according to the literature [31]. The past ten years have seen the publication of a significant number of sparse subset selection-based VS algorithms, however the great majority of them employ a linear sparse formulation in the explicit feature vector space of video frames and disregard either local or global frame associations. This study builds on the conventional sparse subset selection for VS to produce kernel block sparse subset selection by using kernel sparse coding and providing a local inter-frame interaction via packing of frame blocks (KBS3). To enable global inter-frame interaction via similarity, researchers suggest using the SB2S3 technique, which uses similarity-based block sparse subset selection. Finally, a technique for greedy pursuit optimization of NP-hard models is developed. The advantages of SB2S3 include the following: 1) similarity enables an examination of the overall relationship between all frames; 2) block sparse coding considers the local relationship of nearby frames; and 3) it is more broadly applicable because features can generate similarity but not the other way around. Similar to how deep learning-based approaches manage both long-range and short-range dependencies, researchers' model both global and local interactions between frames in this study. Experimental results suggest that this strategy outperforms spare subset selection based and unsupervised deep learning-based techniques of summarization.Long movies are distilled into

easily consumable key frames or key images in a video summary. Recurrent neural networks (RNNs) have recently enabled advancements in supervised methods [32]. Most people make an effort to maximize factual consistency while summarizing. The most basic rule is that they don't care whether the target audience may infer the topic of the video from the synopsis. Existing algorithms are unable to properly maintain summary quality and need huge quantities of training data to avoid overfitting. They assert that creating summaries and inferring the actual information from them constitutes video summarizing. The researchers provide a dual learning architecture that rewards the summary generator with the help of the video reconstructor by integrating summary production (the primary goal) with video reconstruction (the dual task). Two property models are constructed in order to gauge how representative and diversified the summary is. The SumMe, TVsum, OVP, and YouTube datasets, utilized in experiments, show that their technique, which utilizes less training data and compact RNNs as the summary generator, can perform on par with supervised methods that use more complicated summary generators and train on larger datasets of annotated data.

Video summarizing (VS), according to research [33], may help people rapidly understand complicated video information. SDS is a fantastic solution for VS issues when the connection between keyframes and other frames is not precisely linear. This premise is not always valid when working with nonlinear video frames. In this paper, researchers build a nonlinear SDS model for VS using the nonlinearity that exists between video frames. The nonlinearity is converted to linearity by mapping a video into a high-dimensional feature space created by a kernel function. Researchers provide a robust KSDS greedy method with a backtracking technique in addition to a typical kernel SDS (KSDS) greedy approach to handle this problem. A flexible criterion called the energy ratio was developed to provide video summaries of various durations for various video assets. The recommended solution outperforms the top VS algorithms already in use on two common video datasets.A video summary's objective is to provide the most important information from a video in an abbreviated manner. As it stands, dynamic features like camera movement and lighting cannot be handled by current video summary algorithms [34]. This research offers a novel framework for video summarization using eye tracker data since human eyes can reliably follow moving things. The movement of the eyes to follow a moving object is referred to as "smooth chase" in video. Researchers develop a novel technique to distinguish smooth pursuit from fixation and saccade. While smooth pursuit can show you where objects are moving inside a video frame, it is unable to show you whether or not those objects are salient or fascinating to watch. Salient areas and object movements are used to identify key frames for video summarization. In addition, researchers provide a novel spatial saliency prediction approach that builds a saliency map around each smooth pursuit gaze point utilizing elements of the fovea, parafovea, and perifovea of the human visual field. Smooth pursuit is utilized to track motion, and the distance between the viewer's current and prior gaze locations is used to determine the saliency of the motion. Smooth pursuit is motivated by the relationship between visual attention and the apparent velocity of an object. The saliency score for each frame is then combined using the motion and spatial saliency maps, and the most crucial frames are picked using a skimming ratio that the user may choose or that the system will use by default. On video data from an office scenario, replete with several camera angles and lighting conditions, the suggested method's effectiveness is shown. The suggested method of forecasting spatial and motion saliency outperforms (FSMS) conventional best practices, according to experimental findings.

Finding interesting video content is a typical challenge [35]. This is a difficult process since it calls for knowledge with both the whole video and how those parts connect to one another, in addition to the precise target pieces. The Sequence-to-Segments Network is one such innovative end-to-end sequential encoder-decoder system (S2N). S2N encodes the input video into discrete states that record data in real-time. Segments are discovered sequentially by the SDU. The SDU combines the hidden states from the encoder and decoder at each decoding step to identify the target segment. Researchers train the Hungarian Matching Algorithm with Lexicographic Cost to match predicted segments to real data. Researchers advocate using Earth Mover's Distance squared to reduce segment localization errors. They demonstrate that S2N outperforms cutting-edge techniques in the creation of human action recommendations, video summarization, and video highlighting for different use cases.Video summarizing makes an effort to provide a concise summary from several perspectives in order to manage videos on a large scale [36]. The majority of approaches depend on self-attention over numerous video frames, which doesn't properly capture how video frames change. Reviewing pairwise similarity evaluation in the self-attention process has addressed the task of inner-product affinity producing discriminative rather than diverse features. The researchers suggest using squared Euclidean distance to compare people's similarities as one approach. Researchers have created a model of local contextual data that might be used to avoid duplicate video. In order to create SUM-DCA, researchers integrated the two attentional strategies. SUM-F-score DCAs and rank-based algorithms are better, as shown by extensive testing on benchmark datasets.

Work in [37] claims that video synthetic aperture radar may be useful for automated information retrieval and interpretation in dynamic zones of interest (VideoSAR). (DROI). Key-frame extraction enables the processing of huge volumes of video data successfully. Researchers provide a technique for automatically choosing key frames for scattering in VideoSAR using computer vision. A general parameterization model is used to disclose the scattering crucial frames for VideoSAR. Researchers present the sub aperture energy gradient (SEG) and a modified statistical and knowledge-based object tracker to determine the scattering key-frame condition of transient persistence and disappearance (MSAKBOT). With the suggested SEG-MSAKBOT approach, multitemporal video sequences may be efficiently adjusted, enabling more detailed measurements of scattering-critical frames. Finally, the experimental findings and performance assessment on two actuals airborne VideoSAR data with coherent integration angles of 21 ° and 27 ° show the robustness and dependability of the suggested viewpoint to capture scattering key frames and video

content summarization with extremely precise descriptions in various DROIs.Multiview video summary hasn't received much attention from the scholarly community, claims [38]. (MVS). Early MVS studies often ignored fog, depended on summaries, needed more communication capacity and transmission time, and were carried out under less-than-ideal circumstances. An edge intelligence MVS and activity detection system based on the Internet of Things is advised by experts. The architecture of these low-resource devices uses a lightweight object recognition model based on convolutional neural networks (CNN) to split Multiview videos into shots. They developed a system that rapidly and effectively detects activity, computes correlations between camera images, and does so while using less calculation, transmission time, and network capacity levels.

Work in [39] addresses the task of picking pertinent scenes from a movie to include in a synopsis. Modern techniques for summarizing video clips often learn soft importance ratings rather than explicitly using links between the video clips. Consider the interconnection of the source video and choose a selection of clips that have a difficult job associated with them in order to infer a relevant video summary. In this study, researchers identify unsupervised relation-aware hard assignments for choosing significant clips using clip-clip relations. Researchers first construct a network for assignment learning using clips (ClipNet). Based on the size of node properties, scientists draw broad conclusions. The whole framework has been optimized using a multi-task loss that combines a reconstruction and contrastive constraint. Three key metrics have shown the strategy's efficacy.One area that has benefitted from sparse representation is video summaries [40]. (VS). A global description of each frame is used by the great majority of VS algorithms. It is easy to leave out important local information while discussing a worldwide event. This study suggests a technique for splitting video frames into patches and describing them uniformly. Instead, then concatenating the characteristics of each patch and using the standard sparse representation, researchers see each video frame as a block made up of a number of patches. In order to create SBOMP, a simultaneous variation of block-based OMP (Orthogonal Matching Pursuit), researchers take into consideration the reconstruction restriction. The neighbourhood-based approach, which is an extension of the suggested paradigm, sees neighbouring frames as a single super block. One of the first VS block approaches, this one-use blocks in a sparse manner. They ran tests on two popular VS datasets, and the results showed that the suggested methodologies performed as well as or better than supervised deep learning VS methods and state-of-the-art sparse representation-based VS methods.

High-speed and time-lapse videos are often used to show memorable events, despite the fact that various circumstances in the real world have varying motion speeds [41]. After choosing a framerate, each increase or reduction is either undetectable or produces irritating artifacts. Researchers advise standardizing the pixel-space speed of distinct movements in an input video in order to generate a smooth output with spatiotemporally changing framerates (STFR). To recognize and analyze individual movements, researchers advise viewing image-frames many times. Consumers may select the aesthetic constraints that must be adhered to by their designs as a consequence of their creativity. The output that has been motion-normalized effectively shows how a scene changes across different periods.The manual extraction of characteristics used in traditional approaches for creating summaries of Wireless capsule endoscopy (WCE) movies is inadequate to capture the semantic similarity of endoscopic images. For training supervised techniques of extracting deep image features, a lot of annotated data is needed. Utilizing a convolutional auto encoder (CAE), unsupervised features are obtained. Numerous evaluations led to the threshold that is utilized to categorize WCE photos. Keyframe extraction is then used to create a structured WCE video description. The compression ratio and F-measure produced by the suggested approach [42] are 83.12% and 91.1%, respectively. The suggested technique works better than cutting-edge WCE video summarizing techniques.

Studies [43] suggest that the high keyframe recurrence in user-generated videos makes summarizing them difficult. A GAT-modified Bi-LSTM model is offered by the authors for unsupervised video summarization. The GAT first elevates the visual characteristics of an image to a more abstract level using a convolutional feature transform (CFT). A spatial attention model based on salient-area-size is proposed for extracting visual data inside individual frames since individuals often concentrate on large, moving things. The increased degree of integration of visual and semantic input in Bi-LSTM increases the chance of choosing a keyframe. They have the most successful strategy of their type, according to extensive testing sets.The neural network used for the CRSum video description was reported in [44]. The proposed network incorporates three different but connected techniques: feature extraction, temporal modelling, and summarization. The recommended strategy has three unique characteristics: For the first time, it combines trainable three-dimensional convolutional neural networks in combination with feature fusion to create detailed and nuanced video features after using convolutional recurrent neural networks (CNN CRNN) to represent the spatial and temporal structure of videos for summarization. Finally, it adds Sobolev loss I, a novel loss function. Experimental evidence supports the strategy's efficacy. Through carefully planned experiments, scientists may examine the validity of their methods.

Condensing a source text into a manageable format was the focus of research in [45]. Summarization of text is kind of a natural language processing tool. For quick multimedia data delivery over the Internet, multi-modal summarization (MMS) from asynchronous text, image, audio, and video is essential. In this paper, researchers present an extractive MMS technique that combines computer vision, audio processing, and natural language processing for summarizing multimodal news. Semantic gaps in multimodal data should be filled up. The clip combines music and images. The researchers' method enables the selective use of audio transcription and the deduction of salience from audio inputs. Researchers use neural networks to learn both word and image representations. Researchers have shown that multi-modal topic modelling and text-image matching are especially good at capturing the visual content of the

final summary. A textual summary is produced by assessing all multi-modal elements and optimizing submodular functions to maximize salience, non-redundancy, readability, and coverage. MMS corpora are made accessible to researchers in both Chinese and English. The outcomes of their dataset comparison to baseline methods used by rivals show how effective their image-matching and image-topic framework is when evaluated for real-time scenarios.For real-time use cases, video screening to remove objectionable or improper material is typical due to the interactive and social aspect of e-commerce livestreams. It is difficult and time-consuming to analyze multimodal video material, containing video frames and audio samples. To reduce possible hazards, the researchers have created a system called Video Moderator that combines human and machine intelligence. This approach employs powerful machine learning techniques to extract risk-aware video aspects and recognize controversial movies. This framework can support viewpoints from voice, video, and still images. By concentrating on high-risk areas of the video, experts draw attention to potentially unpleasant information. A visual summarization technique called Frame View (FV) [46] was created by academics to blend risk-aware characteristics with the context of videos. The structure of audio view is narrative-based in order to provide users a complete image of the audio material. In addition to a case study, researchers also perform tests and a controlled user study to verify Video Moderator's efficacy.

According to reports, egocentric video is now being studied by the computer vision and multimedia fields [47]. Researchers outline a unified architecture for weakly monitoring the localization, identification, and summary of egocentric video activities at the superpixel level. Researchers' first separate and count the number of events happening in each frame of an egocentric movie before summarizing it. The superpixel level approach improves action localisation and recognition precision. Researchers can extract the superpixels that make up a video frame from the core areas of an egocentric movie using a center-surround paradigm. Superpixels are used in deep feature space to build a sparse spatiotemporal video representation graph (SPRG). Using unsupervised random walks, action labels are created for each superpixel. For each frame, scientists create action labels from superpixels using a fractional knapsack formulation (of actions).Key-shot video summary requires the use of inner-shot and inter-shot interactions [48]. Recurrent neural networks are used in existing techniques to represent video frames. Sequence models neglect higher degree, more distant relationships while emphasizing close-range ties. Although each shot records a discrete activity that changes over time gradually, multi-hop correlations are common. To benefit from this video, you must be aware of both regional and global interdependencies. Consequently, the reconstructive sequence-graph network (RSGN) has been proposed. This network would encode frames and shoots as a sequence and a graph hierarchically, with a long short-term memory (LSTM) capturing relationships at the frame level and a graph convolutional network capturing dependencies at the shot level (GCN). Then, by taking into account both the global and local shot dependencies, the video material is summarized. A reconstructor is developed to reward the summary generator and enable it to get over the absence of annotated data while summarizing videos. The anticipated summary may improve shot-level dependence and video data preservation while reducing reconstruction loss. On three widely-used datasets, their suggested approach has been found to be the most efficient for summarizing tasks (SumMe, TVsum, and VTW).

Vision-based sign language translation (SLT) is a difficult endeavour since sign linguistics uses intricate facial expressions, gestures, and articulated postures. Since SLT is a non-supervised sequence-to-sequence learning problem, its time span is indefinite. Hierarchical deep Recurrent Fusion (HRF) is recommended by professionals in order to comprehend the temporal cues in videos [49]. An adaptive temporal encoder was developed by HRF engineers to record RGB visemes and skeletal signees. Skeletal signees and RGB visemes are learned by ACS. Attention-aware weighting, adaptive temporal pooling, and variable-length clip mining are some of the modules available in ACS. researchers present a query-adaptive decoding fusion to translate the target text based on mismatched behavioural patterns (RGB visemes and skeleton signees). Through extensive testing, the HRF framework has been shown to be successful under different use cases.According to [50], emotional content is crucial for user-generated videos. It is difficult to do a thorough visual emotion analysis of these videos due to the narrow range of face expressions. The authors of this paper recommend using a Bi-stream Emotion Attribution-Classification Network to overcome the difficulties of emotion detection, emotion attribution, and emotion-oriented summarization (BEAC-Net). BEAC-Net has a mechanism in place for tagging and categorizing data. A dataset's most significant emotional component may be found using attribution networks, which can then be used to categorize the dataset and reduce its sparsity. An excerpt and the original video are both fed into a bi-stream classification network. Researchers provide a novel dataset containing ground-truth labels for emotional states that have been annotated by humans. Experimental results on two video datasets demonstrate the framework's improved performance and the complimentary nature of the dual classification streams. But these models are highly variant in terms of their quantitative performance, thus, an empirical analysis of these models is discussed in the next section of this paper.

3. Statistical Analysis

Based on the in-depth theoretical review on summarization models, it can be observed that existing models are considerably variant in terms of their functional characteristics. This makes it difficult for researchers to identify optimal models for their performance-specific deployments. To simplify this task, an empirical analysis of the reviewed summarization models in terms of summarization accuracy (A), delay (D), complexity (C), scalability (S) and fMeasure (F) is discussed in this section, which will assist readers to identify optimal models as per their deployment

requirements. Values of Accuracy & fMeasure were directly evaluated from the reviewed models, but, delay, complexity & scalability were not available directly in terms of absolute value metrics. Thus, these value metrics were quantized into fuzzy levels of, Low (L=1), Medium (M=2), High (H=3), and Very High (VH=4) ranges, depending upon their internal implementation characteristics. Due to these evaluations, readers will be able to identify optimum models for different use cases (models for high accuracy, low delay, etc.), thus covering objectives of this research under real-time deployments. For instance, CNNs have higher complexity than linear classification models, thus, their complexity fuzzy levels will be higher, which will assist readers in identifying optimal models as per individual metric sets. All the data present in table 1 is taken directly from the respective reference papers, thus readers can easily identify performance characteristics for the compared models and use this data for their use cases. Based on this strategy, the empirical evaluation of these metrics is depicted in table 1 as follows,

Table 1. Empirical analysis of different summarization models for video sequence sets

Method	Accuracy (%)	Delay	Complexity	fMeasure	Scalability (S)
SA [1]	90.5	M	VH	88.53	VH
3DST-UNet-RL [2]	98.6	VH	H	89.91	H
RL [3]	94.2	H	VH	85.81	H
SIFT CNN [4]	94.9	H	H	85.34	VH
Elastic VS [5]	85.5	VH	H	80.53	H
SMK [6]	92.7	M	VH	82.44	L
SDSOF AM [7]	79.5	H	H	82.50	VH
CCA LSTM [8]	91.6	H	VH	86.88	H
FCNLectureNet [9]	92.9	VH	H	86.44	H
BiLSTM [10]	93.5	VH	H	83.50	VH
DSSE [11]	90.2	H	VH	83.00	M
DNMVS [12]	83.5	H	H	84.88	H
CNN HSTM [13]	91.9	VH	VH	87.06	VH
Deep CNN [14]	96.2	VH	H	89.25	VH
ADSum [15]	90.5	H	M	82.81	M
ACSUMGAN [16]	98.9	VH	H	79.38	M
QSAN [17]	75.6	H	VH	75.13	L
DCNN [18]	79.5	VH	H	81.47	H
ASVS [19]	85.3	H	H	85.00	L
DB-LSTM [20]	95.9	VH	VH	81.72	H
Dual LSTM [21]	90.8	VH	H	80.31	VH
Emotion Net [22]	74.8	H	H	81.28	L
MetaL-TDVS [23]	91.4	VH	VH	78.34	M
TTH-RNN [24]	93.9	H	H	75.81	H
KF [25]	65.4	H	H	74.44	L
DSNet [26]	83.3	M	H	82.25	H
DSSR2 [27]	89.5	H	H	80.75	VH
CNN [28]	90.4	H	M	82.88	VH
Char KF [29]	78.5	VH	H	83.19	H
GCNLSTM Interaction [30]	96.3	VH	H	88.00	H
SB2S3 [31]	91.4	H	H	81.50	H
Dual RNN [32]	93.9	VH	H	79.03	H
Kernel SDS [33]	75.5	H	H	77.53	VH
FSMS [34]	83.5	H	M	82.13	H
S2N [35]	89.1	M	H	86.38	H
SUMDCA [36]	90.2	H	H	86.81	H
SEGMSAKBOT [37]	97.1	H	VH	86.59	VH
MVS CNN [38]	90.5	H	H	79.81	H
ClipNet [39]	89.5	VH	H	79.81	L
SBOMP [40]	75.4	M	H	80.94	H
STFR [41]	90.5	H	H	87.53	H
CAE [42]	93.1	VH	H	88.28	H
GAT Bi-LSTM [43]	96.5	H	H	88.63	H
CNN CRNN [44]	92.9	M	M	85.19	VH
MMS CNN [45]	94.2	H	H	82.25	H
FV [46]	85.5	H	VH	77.78	H
SPRG [47]	83.5	VH	H	79.34	VH
RSGN [48]	79.9	H	VH	83.13	H
HRF [49]	90.5	H	H	85.18	M
BEACNet [50]	95.6	VH	H	83.68	H

Based on the results of this thorough analysis, it can be concluded that the applications for high-efficiency video summarization include ACSUM GAN [16], 3DST-UNet-RL [2], SEG MSAK BOT [37], GAT Bi-LSTM [43], GCN LSTM Interaction [30], Deep CNN [14], DB-LSTM [20], BEAC Net [50], and SIFT CNN [4]. These systems also demonstrate the highest accuracy. When multiple video types are available for analysis, these models are helpful because they use deep learning techniques, which increase their accuracy. The fastest systems, SA [1], SMK [6], DSNet

[26], S2N [35], SBOMP [40], and CNN CRNN [44], can be used for low-delay summarization use cases. The use of parallel processing and pipelining during the processing of videos is the reason for this efficiency optimization because these techniques are very effective for complex operations. As a result of their lower complexity and suitability for applications with constrained computational resources, the models suggested in ADSum [15], CNN [28], FSMS [34], and CNN CRNN [44] have been adopted. This results from the feature extraction and processing operations using streamlined computational layers. A higher fMeasure is displayed by the models proposed in 3DST-UNet-RL [2], Deep CNN [14], GAT Bi-LSTM [43], SA [1], CAE [42], GCN LSTM Interaction [30], STFR [41], and CNN HSTM [13], which makes them useful for high precision and high recall applications. The models discussed in SA [1], SIFT CNN [4], SDSOF AM [7], BiLSTM [10], CNN HSTM [13], Deep CNN [14], Dual LSTM [21], DSSR2 [27], CNN [28], Kernel SDS [33], SEG MSAK BOT [37], CNN CRNN [44], and SPRG [47] can be combined with this performance to increase their scalability levels and make them useful for multiple domain video summarization applicationsTo further facilitate model selection, these metrics were combined to evaluate a novel Summarization Rank Metric (SRM) which is calculated via equation 1,

$$\text{SRM} = \frac{\text{A+F}}{200} + \frac{\text{S}}{4} + \frac{1}{\text{D}} + \frac{1}{\text{C}} \tag{1}$$

Based on this evaluation, and figure 1, it can be observed that models proposed in CNN CRNN [44], CNN [28], SA [1], SIFT CNN [4], DSSR2 [27], Deep CNN [14], SEG MSAK BOT [37], SDSOF AM [7], BiLSTM [10], S2N [35], Dual LSTM [21], Kernel SDS [33], FSMS [34], and DSNet [26] showcase overall highest performance, which makes them applicable for high accuracy, high fMeasure, high scalability, low delay and low complexity scenarios.

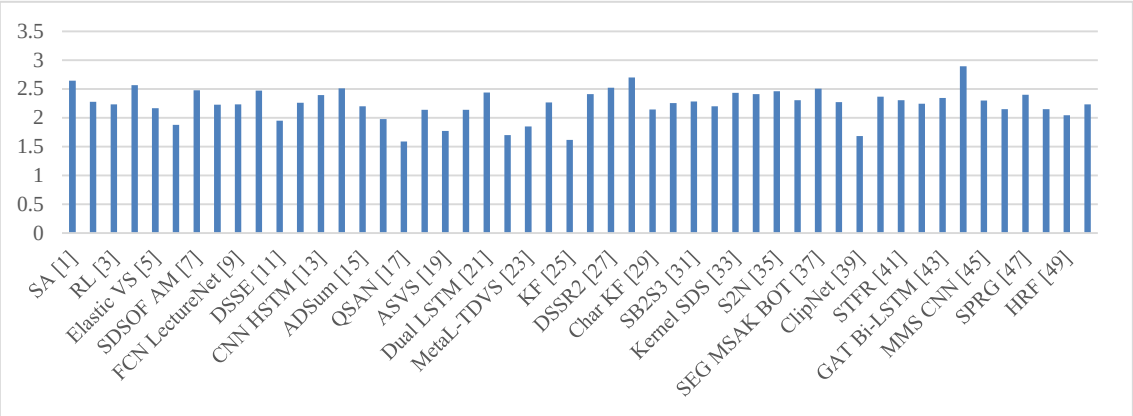

Fig. 1. SRM for different summarization models

Thus, it is recommended that researchers must use these models and their extensions for high-performance video summarization deployments.

4. Conclusion & Future Scope

This paper performs a detailed analysis of different summarization models in terms of their functional nuances, application-specific advantages, deployment-specific limitations, and contextual future scopes. Based on this discussion it was observed that deep learning models outperform others in terms of their accuracy, precision, recall, and other quantitative measures, which makes them highly useful for summarization deployments. It was also observed that linear models showcase moderate performance, which limits their applicability & scalability performance levels. On the basis of empirical research conducted on these models, it was discovered that ACSUM GAN, DST-UNet-RL, SEG MSAK BOT, GAT Bi-LSTM, GCN LSTM Interaction, Deep CNN, DB-LSTM, BEAC Net, and SIFT CNN showcase the highest accuracy and, as a result, can be utilized in high-efficiency video summarization applications. These models are also helpful in situations where there are a variety of video types available for analysis. It has also been observed that SA, SMK, DSNet, SN, and SBOMP showcase the highest speed, and as a result, they are candidates for use in CNN CRNN use cases that require low-delay summarization. The models that were proposed in ADSum, CNN, FSMS, and CNN CRNN all have a lower level of complexity, and as a result, they are suitable for use in applications that have restricted computational capabilities. While other models, such as those proposed in DST-UNet-RL, Deep CNN, GAT Bi-LSTM, SA, CAE, GCN LSTM Interaction, STFR, and CNN HSTM, exhibit higher fMeasure, which enables these models to be useful for applications that require both high precision and high recall. This performance can be combined with models that have been discussed in SA, SIFT CNN, SDSOF AM, BiLSTM, CNN HSTM, Deep CNN, Dual LSTM, DSSR, CNN, Kernel SDS, SEG MSAK BOT, CNN CRNN, and SPRG. This will assist in improving the scalability levels of those models and make them useful for multiple domain video summarization applications. It was discovered that models proposed in CNN CRNN, CNN, SA, SIFT CNN, DSSR, Deep CNN, SEG MSAK BOT, SDSOF AM,

BiLSTM, SN, Dual LSTM, Kernel SDS, FSMS, and DSNet showcase the overall highest performance, which enables them to be applicable for high accuracy, high fMeasure, high scalability, low delay, and low complexity scenarios. This was accomplished by combining these metrics to Because of this, it is strongly suggested that researchers make use of these models and the extensions that they have created for high-performance video summarization deployments. In future, researchers can extend these models via integration of hybrid bioinspired models for keyframe extraction, and identification of redundancies in video sequences. Moreover, validation of these models on larger sets can be done, and their performance can be improved via fusion of multiple deep learning models, which will assist in better feature representation & analysis capabilities.

References

[1] C. Huang and H. Wang, "A Novel Key-Frames Selection Framework for Comprehensive Video Summarization," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 2, pp. 577-589, Feb. 2020, doi: 10.1109/TCSVT.2019.2890899.

[2] T. Liu, Q. Meng, J. -J. Huang, A. Vlontzos, D. Rueckert and B. Kainz, "Video Summarization Through Reinforcement Learning With a 3D Spatio-Temporal U-Net," in IEEE Transactions on Image Processing, vol. 31, pp. 1573-1586, 2022, doi: 10.1109/TIP.2022.3143699.

[3] J. Lei, Q. Luan, X. Song, X. Liu, D. Tao and M. Song, "Action Parsing-Driven Video Summarization Based on Reinforcement Learning," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 7, pp. 2126-2137, July 2019, doi: 10.1109/TCSVT.2018.2860797.

[4] O. Issa and T. Shanableh, "CNN and HEVC Video Coding Features for Static Video Summarization," in IEEE Access, vol. 10, pp. 72080-72091, 2022, doi: 10.1109/ACCESS.2022.3188638.

[5] Y. Wang, Y. Dong, S. Guo, Y. Yang and X. Liao, "Latency-Aware Adaptive Video Summarization for Mobile Edge Clouds," in IEEE Transactions on Multimedia, vol. 22, no. 5, pp. 1193-1207, May 2020, doi: 10.1109/TMM.2019.2939753.

[6] P. D. Byrnes and W. E. Higgins, "Efficient Bronchoscopic Video Summarization," in IEEE Transactions on Biomedical Engineering, vol. 66, no. 3, pp. 848-863, March 2019, doi: 10.1109/TBME.2018.2859322.

[7] S. S. Thomas, S. Gupta and V. K. Subramanian, "Context Driven Optimized Perceptual Video Summarization and Retrieval," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 10, pp. 3132-3145, Oct. 2019, doi: 10.1109/TCSVT.2018.2873185.

[8] L. Yuan, F. E. H. Tay, P. Li and J. Feng, "Unsupervised Video Summarization With Cycle-Consistent Adversarial LSTM Networks," in IEEE Transactions on Multimedia, vol. 22, no. 10, pp. 2711-2722, Oct. 2020, doi: 10.1109/TMM.2019.2959451.

[9] K. Davila, F. Xu, S. Setlur and V. Govindaraju, "FCN-LectureNet: Extractive Summarization of Whiteboard and Chalkboard Lecture Videos," in IEEE Access, vol. 9, pp. 104469-104484, 2021, doi: 10.1109/ACCESS.2021.3099427.

[10] Z. Ji, K. Xiong, Y. Pang and X. Li, "Video Summarization With Attention-Based Encoder–Decoder Networks," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 6, pp. 1709-1717, June 2020, doi: 10.1109/TCSVT.2019.2904996.

[11] Y. Yuan, T. Mei, P. Cui and W. Zhu, "Video Summarization by Learning Deep Side Semantic Embedding," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 1, pp. 226-237, Jan. 2019, doi: 10.1109/TCSVT.2017.2771247.

[12] W. Zhu, Y. Han, J. Lu and J. Zhou, "Relational Reasoning Over Spatial-Temporal Graphs for Video Summarization," in IEEE Transactions on Image Processing, vol. 31, pp. 3017-3031, 2022, doi: 10.1109/TIP.2022.3163855.

[13] S. Huang, X. Li, Z. Zhang, F. Wu and J. Han, "User-Ranking Video Summarization With Multi-Stage Spatio–Temporal Representation," in IEEE Transactions on Image Processing, vol. 28, no. 6, pp. 2654-2664, June 2019, doi: 10.1109/TIP.2018.2889265.

[14] M. Ma et al., "Keyframe Extraction From Laparoscopic Videos via Diverse and Weighted Dictionary Selection," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 5, pp. 1686-1698, May 2021, doi: 10.1109/JBHI.2020.3019198.

[15] Z. Ji, Y. Zhao, Y. Pang, X. Li and J. Han, "Deep Attentive Video Summarization With Distribution Consistency Learning," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 4, pp. 1765-1775, April 2021, doi: 10.1109/TNNLS.2020.2991083.

[16] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris and I. Patras, "AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 8, pp. 3278-3292, Aug. 2021, doi: 10.1109/TCSVT.2020.3037883.

[17] S. Xiao, Z. Zhao, Z. Zhang, Z. Guan and D. Cai, "Query-Biased Self-Attentive Network for Query-Focused Video Summarization," in IEEE Transactions on Image Processing, vol. 29, pp. 5889-5899, 2020, doi: 10.1109/TIP.2020.2985868.

[18] K. Muhammad, T. Hussain, M. Tanveer, G. Sannino and V. H. C. de Albuquerque, "Cost-Effective Video Summarization Using Deep CNN With Hierarchical Weighted Fusion for IoT Surveillance Networks," in IEEE Internet of Things Journal, vol. 7, no. 5, pp. 4455-4463, May 2020, doi: 10.1109/JIOT.2019.2950469.

[19] A. Dilawari and M. U. G. Khan, "ASoVS: Abstractive Summarization of Video Sequences," in IEEE Access, vol. 7, pp. 29253-29263, 2019, doi: 10.1109/ACCESS.2019.2902507.

[20] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik and V. H. C. de Albuquerque, "Cloud-Assisted Multiview Video Summarization Using CNN and Bidirectional LSTM," in IEEE Transactions on Industrial Informatics, vol. 16, no. 1, pp. 77-86, Jan. 2020, doi: 10.1109/TII.2019.2929228.

[21] Z. Zhang, D. Xu, W. Ouyang and C. Tan, "Show, Tell and Summarize: Dense Video Captioning Using Visual Cue Aided Sentence Summarization," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 9, pp. 3130-3139, Sept. 2020, doi: 10.1109/TCSVT.2019.2936526.

[22] H. Zeng et al., "EmotionCues: Emotion-Oriented Visual Summarization of Classroom Videos," in IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 7, pp. 3168-3181, 1 July 2021, doi: 10.1109/TVCG.2019.2963659.

[23] X. Li, H. Li and Y. Dong, "Meta Learning for Task-Driven Video Summarization," in IEEE Transactions on Industrial Electronics, vol. 67, no. 7, pp. 5778-5786, July 2020, doi: 10.1109/TIE.2019.2931283.
[24] B. Zhao, X. Li and X. Lu, "TTH-RNN: Tensor-Train Hierarchical Recurrent Neural Network for Video Summarization," in IEEE Transactions on Industrial Electronics, vol. 68, no. 4, pp. 3629-3637, April 2021, doi: 10.1109/TIE.2020.2979573.
[25] K. Muhammad, T. Hussain, J. Del Ser, V. Palade and V. H. C. de Albuquerque, "DeepReS: A Deep Learning-Based Video Summarization Strategy for Resource-Constrained Industrial Surveillance Scenarios," in IEEE Transactions on Industrial Informatics, vol. 16, no. 9, pp. 5938-5947, Sept. 2020, doi: 10.1109/TII.2019.2960536.
[26] W. Zhu, J. Lu, J. Li and J. Zhou, "DSNet: A Flexible Detect-to-Summarize Network for Video Summarization," in IEEE Transactions on Image Processing, vol. 30, pp. 948-962, 2021, doi: 10.1109/TIP.2020.3039886.
[27] M. Ma, S. Mei, S. Wan, Z. Wang, X. -S. Hua and D. D. Feng, "Graph Convolutional Dictionary Selection With L_2 ,□ Norm for Video Summarization," in IEEE Transactions on Image Processing, vol. 31, pp. 1789-1804, 2022, doi: 10.1109/TIP.2022.3146012.
[28] T. Hussain, K. Muhammad, J. D. Ser, S. W. Baik and V. H. C. de Albuquerque, "Intelligent Embedded Vision for Summarization of Multiview Videos in IIoT," in IEEE Transactions on Industrial Informatics, vol. 16, no. 4, pp. 2592-2602, April 2020, doi: 10.1109/TII.2019.2937905.
[29] P. Zhou et al., "Character-Oriented Video Summarization With Visual and Textual Cues," in IEEE Transactions on Multimedia, vol. 22, no. 10, pp. 2684-2697, Oct. 2020, doi: 10.1109/TMM.2019.2960594.
[30] Z. Zhang, D. Xu, W. Ouyang and L. Zhou, "Dense Video Captioning Using Graph-Based Sentence Summarization," in IEEE Transactions on Multimedia, vol. 23, pp. 1799-1810, 2021, doi: 10.1109/TMM.2020.3003592.
[31] M. Ma, S. Mei, S. Wan, Z. Wang, D. D. Feng and M. Bennamoun, "Similarity Based Block Sparse Subset Selection for Video Summarization," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 10, pp. 3967-3980, Oct. 2021, doi: 10.1109/TCSVT.2020.3044600.
[32] B. Zhao, X. Li and X. Lu, "Property-Constrained Dual Learning for Video Summarization," in IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 10, pp. 3989-4000, Oct. 2020, doi: 10.1109/TNNLS.2019.2951680.
[33] M. Ma, S. Mei, S. Wan, Z. Wang and D. Feng, "Video Summarization via Nonlinear Sparse Dictionary Selection," in IEEE Access, vol. 7, pp. 11763-11774, 2019, doi: 10.1109/ACCESS.2019.2891834.
[34] M. Paul and M. MusfequsSalehin, "Spatial and Motion Saliency Prediction Method Using Eye Tracker Data for Video Summarization," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 6, pp. 1856-1867, June 2019, doi: 10.1109/TCSVT.2018.2844780.
[35] Z. Wei et al., "Sequence-to-Segments Networks for Detecting Segments in Videos," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 3, pp. 1009-1021, 1 March 2021, doi: 10.1109/TPAMI.2019.2940225.
[36] Y. Pan et al., "Exploring Global Diversity and Local Context for Video Summarization," in IEEE Access, vol. 10, pp. 43611-43622, 2022, doi: 10.1109/ACCESS.2022.3163414.
[37] Y. Zhang, D. Zhu, H. Bi, G. Zhang and H. Leung, "Scattering Key-Frame Extraction for Comprehensive VideoSAR Summarization: A Spatiotemporal Background Subtraction Perspective," in IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 7, pp. 4768-4784, July 2020, doi: 10.1109/TIM.2019.2953435.
[38] T. Hussain et al., "Multiview Summarization and Activity Recognition Meet Edge Computing in IoT Environments," in IEEE Internet of Things Journal, vol. 8, no. 12, pp. 9634-9644, 15 June15, 2021, doi: 10.1109/JIOT.2020.3027483.
[39] J. Gao, X. Yang, Y. Zhang and C. Xu, "Unsupervised Video Summarization via Relation-Aware Assignment Learning," in IEEE Transactions on Multimedia, vol. 23, pp. 3203-3214, 2021, doi: 10.1109/TMM.2020.3021980.
[40] S. Mei, M. Ma, S. Wan, J. Hou, Z. Wang and D. D. Feng, "Patch Based Video Summarization With Block Sparse Representation," in IEEE Transactions on Multimedia, vol. 23, pp. 732-747, 2021, doi: 10.1109/TMM.2020.2987683.
[41] S. Wehrwein, K. Bala and N. Snavely, "Scene Summarization via Motion Normalization," in IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 4, pp. 2495-2501, 1 April 2021, doi: 10.1109/TVCG.2020.2993195.
[42] B. Sushma and P. Aparna, "Summarization of Wireless Capsule Endoscopy Video Using Deep Feature Matching and Motion Analysis," in IEEE Access, vol. 9, pp. 13691-13703, 2021, doi: 10.1109/ACCESS.2020.3044759.
[43] R. Zhong, R. Wang, Y. Zou, Z. Hong and M. Hu, "Graph Attention Networks Adjusted Bi-LSTM for Video Summarization," in IEEE Signal Processing Letters, vol. 28, pp. 663-667, 2021, doi: 10.1109/LSP.2021.3066349.
[44] Y. Yuan, H. Li and Q. Wang, "Spatiotemporal Modeling for Video Summarization Using Convolutional Recurrent Neural Network," in IEEE Access, vol. 7, pp. 64676-64685, 2019, doi: 10.1109/ACCESS.2019.2916989.
[45] H. Li, J. Zhu, C. Ma, J. Zhang and C. Zong, "Read, Watch, Listen, and Summarize: Multi-Modal Summarization for Asynchronous Text, Image, Audio and Video," in IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 5, pp. 996-1009, 1 May 2019, doi: 10.1109/TKDE.2018.2848260.
[46] T. Tang, Y. Wu, Y. Wu, L. Yu and Y. Li, "VideoModerator: A Risk-aware Framework for Multimodal Video Moderation in E-Commerce," in IEEE Transactions on Visualization and Computer Graphics, vol. 28, no. 1, pp. 846-856, Jan. 2022, doi: 10.1109/TVCG.2021.3114781.
[47] A. Sahu and A. S. Chowdhury, "Together Recognizing, Localizing and Summarizing Actions in Egocentric Videos," in IEEE Transactions on Image Processing, vol. 30, pp. 4330-4340, 2021, doi: 10.1109/TIP.2021.3070732.
[48] B. Zhao, H. Li, X. Lu and X. Li, "Reconstructive Sequence-Graph Network for Video Summarization," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 5, pp. 2793-2801, 1 May 2022, doi: 10.1109/TPAMI.2021.3072117.
[49] D. Guo, W. Zhou, A. Li, H. Li and M. Wang, "Hierarchical Recurrent Deep Fusion Using Adaptive Clip Summarization for Sign Language Translation," in IEEE Transactions on Image Processing, vol. 29, pp. 1575-1590, 2020, doi: 10.1109/TIP.2019.2941267.
[50] G. Tu, Y. Fu, B. Li, J. Gao, Y. Jiang and X. Xue, "A Multi-Task Neural Approach for Emotion Attribution, Classification, and Summarization," in IEEE Transactions on Multimedia, vol. 22, no. 1, pp. 148-159, Jan. 2020, doi: 10.1109/TMM.2019.2922129.

Authors' Profiles

DarshankumarBillur received the Bachelor of Electronics & Communication Engineering from VEC Bellary. Obtained his masters from Basaveshwara Engineering College Bagalkot. Currently pursuing Ph.D in the field of video summarization from Visvesvaraya Technological University, Belgaum. He is interested in field of deep learning and Embedded systems, Cloud computing systems.

Dr. Manu T.M. received the Bachelor of Electronics & Communication Engineering from B.D.T. College of Engineering, Davanagere. Obtained his masters from B.V.Bhoomreddy College of Enggg., Hubli. He obtained his Doctorate from Visvesvaraya Technological University, Belgaum.
He is actively involved in grooming young researchers for achieving significant contribution in their field of interest. He is an author of a great deal of research studies published at national and international journals as well as conference proceedings.

Vishwas Patil graduated from AITM, Batkal in Electronics & Communication Engineering and completed his post graduation from NITK, Suratkal. He is currently persuing his PhD from Visvesvaraya Technological University, Belgaum university. His areas of interest are Signal Processing and VLSI.

How to cite this paper: Darshankumar D.Billur, Manu T. M., Vishwas Patil, "A Comparative Analysis of Video Summarization Techniques", International Journal of Engineering and Manufacturing (IJEM), Vol.13, No.3, pp. 10-24, 2023. DOI:10.5815/ijem.2023.03.02