IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125186

Video Summarization and Translation into Indian Regional Languages Using Deep Learning

Dr Ranjit KN[1], Mr. Ganesh Nayak R[2], Ms. Monica M[3], Ms. Poornima R[4]

[1]Professor, Dept. of Computer Science and Engineering, Maharaja Institute of Technology, Thandavapura

[2345]Students, Dept of Computer Science and Engineering, Maharaja Institute of Technology, Thandavapura

Abstract: The rapid expansion of video content across platforms like YouTube has brought with it challenges related to information overload and accessibility. Many users find it difficult to extract meaningful insights from long videos, while language remains a major barrier for non-English speakers in India. This paper presents a dual-solution system that automatically summarizes video content using an LSTM-based deep learning model and then translates the summary into several Indian regional languages using Google Translate. The system is designed for efficient information access and inclusive communication, ensuring that users can comprehend content quickly and in their preferred language.

Keywords: Video Summarization, LSTM, Natural Language Processing, Deep Learning, Google Translate, Accessibility, Speech Recognition.

I. INTRODUCTION

In the digital era, video content has become one of the most dominant and widely consumed forms of media, particularly in sectors such as education, news, tutorials, and public communication. As the volume and length of videos continue to grow, users face an increasing challenge in extracting relevant information efficiently. This challenge is especially evident in countries such as India, which has a highly diverse and multilingual population. While English is often the primary language for digital content, a significant portion of the population, particularly in rural and non-metropolitan areas, prefers regional languages for better understanding and accessibility. The over-reliance on English-only content creates a digital divide, hindering many people from engaging with valuable educational or informational videos, thus limiting opportunities for learning and growth.

To address these growing challenges, there is a clear need for a tool that condenses long-form video content into concise, relevant summaries and makes this summarized content available in multiple languages. The proposed system combines two powerful technologies to solve this problem: video summarization and multilingual translation. The video summarization component uses advanced techniques, such as LSTM (Long Short-Term Memory) networks, to extract the most important information from lengthy videos, creating a condensed version that retains the core message. This allows users to quickly access the essential content without having to watch the entire video, saving both time and effort.

The multilingual translation feature further enhances the system's utility by providing translated summaries in a variety of regional languages, such as Hindi, Bengali, Tamil, Marathi, and others. This aspect ensures that the content becomes accessible to a wider demographic, overcoming the language barrier and promoting inclusivity. By integrating video summarization and translation into a single, streamlined pipeline, this system empowers users to access, understand, and engage with video content more effectively, regardless of their language or location. Ultimately, it aims to improve accessibility, foster knowledge sharing, and bridge the digital divide in a rapidly evolving world.

II. PROBLEM STATEMENT AND OBJECTIVE

A. Problem statement:

In today's digital age, video content has become a dominant medium for learning, entertainment, and information dissemination. However, users often face the challenge of navigating through long videos to find specific or relevant information. This leads to time inefficiency, cognitive fatigue, and reduced engagement. The inability to quickly extract meaningful insights from extensive video content creates a barrier, especially for students, professionals, and researchers who rely heavily on time-managed learning.

Moreover, a significant portion of online educational and informational video content is available only in English. This presents a language barrier for non-English speakers, particularly in a linguistically diverse country like India. Millions of individuals are excluded from accessing valuable knowledge due to the lack of support for regional

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125186

languages. As a result, even summarized content—if it remains in English—fails to bridge the accessibility gap. Addressing these issues requires a system capable of both summarizing long videos and making that summary linguistically inclusive. Therefore, this project aims to tackle the dual challenge of video summarization and multilingual accessibility.

B. Objective:

The primary objective of this project is to develop an intelligent, end-to-end system that significantly enhances the accessibility and understanding of video content for a diverse user base. With the ever-increasing volume of online video content, especially on platforms like YouTube, users often struggle to consume lengthy videos or find specific information efficiently. This system addresses that challenge by enabling users to obtain concise, meaningful summaries of video content in both English and various Indian regional languages.

The process begins with the user providing a YouTube video link through an easy-to-use interface. The system then automatically downloads and extracts the audio stream from the video. This audio is processed using advanced speech-to-text technology, which transcribes the spoken content into English text. The transcription engine is built to handle different speech patterns, accents, and varying audio qualities, aiming to generate an accurate textual representation of the video dialogue.

Once the transcript is generated, it undergoes cleaning and preprocessing to remove irrelevant elements such as filler words, false starts, and background noise artifacts. The cleaned transcript is then fed into a deep learning-based NLP model, specifically a Long Short-Term Memory (LSTM) network, which is capable of understanding contextual relationships in sequential data. The LSTM model generates a coherent and concise summary that captures the core message of the video content.

To extend the reach of the system, the English summary is translated into multiple Indian languages using a reliable translation API, such as Google Translate. This multilingual output ensures that users from different linguistic backgrounds can comprehend the content easily. The final output is presented through a user-friendly interface, displaying both the English and translated summaries side by side. By combining summarization with multilingual support, this system reduces the time and effort needed to understand long videos and promotes inclusivity by overcoming language barriers.

III. RELATED WORK

Video summarization has been an active research area for more than two decades, evolving from basic keyframe extraction techniques to sophisticated models leveraging deep learning. Early summarization systems were primarily based on visual cues, where algorithms would select representative frames or scenes based on changes in color, motion, or shot boundaries. These systems, while simple, often lacked an understanding of the semantic content and failed to produce coherent summaries for spoken or narrative-heavy videos.

With the advent of machine learning, particularly unsupervised and supervised learning methods, summarization began to incorporate temporal analysis and user-driven objectives. Techniques such as clustering and Hidden Markov Models (HMMs) were applied to detect patterns in frame sequences or transitions, which provided moderate improvements but still lacked contextual depth.

The rise of deep learning brought about a significant shift. Long Short-Term Memory (LSTM) networks and other Recurrent Neural Networks (RNNs) enabled models to process video or audio transcripts as sequences, allowing better handling of contextual dependencies. One key advancement was the introduction of encoder-decoder architectures for text summarization, where the encoder captures the context of the transcript, and the decoder generates a shorter version while maintaining semantic meaning. This architecture inspired many modern video summarization systems that process speech transcripts instead of relying solely on visual content.

Recent works have explored the use of reinforcement learning in summarization. For instance, Deep Summarization Networks (DSN) and frameworks that reward diversity and representativeness have shown promising results in generating user-adapted summaries. These systems treat video summarization as a sequential decision-making process, optimizing summary quality through trial and error. Another notable contribution is the use of adversarial learning with LSTM networks, where a generator creates a summary and a discriminator evaluates its authenticity against human summaries.

Parallel to these advancements in summarization, natural language translation has evolved rapidly. Neural Machine Translation (NMT) models, particularly those based on attention mechanisms and transformers, have outperformed traditional rule-based or statistical approaches. Tools like Google Translate now use these models to provide relatively accurate translations across over 100 languages. However, real-time integration of summarization with multilingual translation remains limited, especially in user-facing applications.

Despite these advances, few systems attempt to combine summarization and translation in a cohesive pipeline. Most translation-focused tools rely on already available text and do not consider the summarization of content as a

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125186

preprocessing step. Likewise, summarization models often focus on English-only datasets, neglecting the linguistic diversity of audiences, particularly in multilingual countries like India.

In contrast, the system proposed in this paper uniquely combines LSTM-based summarization of YouTube video transcripts with translation support for multiple Indian regional languages. By addressing both summarization and language accessibility in a single framework, it caters to a broader and more diverse user base. The approach also leverages the strengths of mature technologies like Google Translate while incorporating customized summarization models that are trained on speech-derived text, providing a balance between innovation and practicality.

IV. SYSTEM DESIGN

The system consists of a multi-stage process that begins with user input and ends with the delivery of a summarized and translated output. It is built to be modular, so each component—from transcription to translation—can be updated or improved independently.

A. Key Components

The proposed system is a comprehensive pipeline designed to transform raw video content into concise, multilingual summaries, streamlining the process of video summarization and translation. The process begins with the User Input module, where users submit a YouTube video link through an intuitive interface. This dynamic entry point allows the system to fetch and process video content without requiring manual uploads, making it user-friendly and seamless.

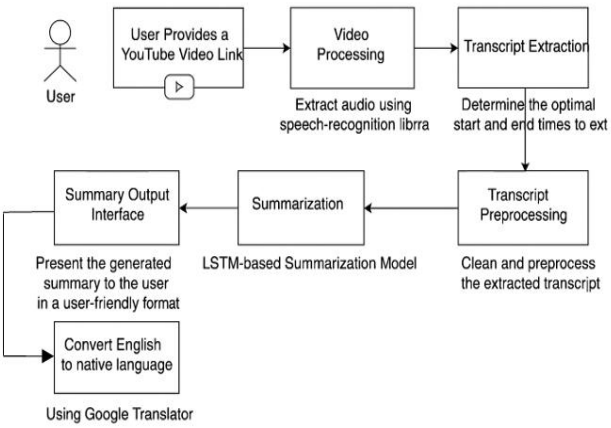
Once the video is fetched, the system proceeds to the Audio Extraction module, where it isolates the audio stream from the video. This step ensures that only the spoken content is extracted, removing any irrelevant background noise, visual elements, or multimedia features, which are unnecessary for textual summarization. The extracted audio is then passed to the Speech-to-Text Transcription module. Here, an Automatic Speech Recognition (ASR) engine converts the audio into a structured transcript. This module is crucial as the accuracy of transcription directly influences the quality of the final summary. The ASR engine is designed to handle various accents, intonations, and domain-specific vocabulary, ensuring the transcription is as accurate as possible.

After the transcription is completed, the system enters the Transcript Preprocessing phase. During this step, several cleaning tasks are carried out, such as removing filler words (e.g., "um," "like"), correcting grammatical errors, standardizing punctuation, and segmenting sentences. These tasks prepare the transcript for summarization by enhancing its linguistic and syntactical quality, ensuring that only meaningful linguistic units are used for the next phase.

The refined transcript is then passed to the Summarization module, which uses a trained Long Short-Term Memory (LSTM) model to generate a concise summary. The LSTM model is designed to understand the sequential context of the text, identifying the most relevant information and extracting it to form a coherent summary that captures the essence of the original video content.

Following summarization, the Translation module translates the English summary into the user's preferred Indian language using the Google Translate API. It supports several regional languages, including Hindi, Tamil, Telugu, Bengali, Marathi, Kannada, Gujarati, and Malayalam, ensuring the summary is accessible to a diverse audience while retaining linguistic accuracy and cultural nuances.

Figure 1: System Architecture for Video Summarization



IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125186

Finally, the Output Interface presents both the English summary and the translated version side by side. This allows users to cross-reference and fully understand the content in their native language. The system's end-to-end design offers an efficient, accessible, and user-centric solution for summarizing and translating video content.

V. METHODOLOGY

The methodology adopted for this system integrates deep learning-based sequence modeling, natural language processing, and real-time API-based translation. The workflow begins with a transcript extracted from a YouTube video, which is then preprocessed, summarized using a trained LSTM model, and translated into the user's selected Indian language. Each stage is crucial to ensure semantic preservation, grammatical accuracy, and user accessibility.

A. LSTM-Based Summarization

Long Short-Term Memory (LSTM) networks are a specialized type of Recurrent Neural Network (RNN) introduced to address the limitations of traditional RNNs in learning long-term dependencies. Unlike simple RNNs, which suffer from vanishing and exploding gradient problems, LSTMs incorporate memory cells and gating mechanisms—specifically the input gate, forget gate, and output gate—to manage the flow of information across time steps. This allows LSTMs to maintain relevant information over longer textual sequences, making them well-suited for tasks such as text summarization, where understanding context and narrative flow is critical.

In this system, the LSTM is employed to generate extractive and partially abstractive summaries of video transcripts. The model is trained on a corpus of manually curated and annotated video transcript summaries. During training, the network learns to recognize patterns and determine which sentences or segments contribute most to the core meaning of the transcript. The network uses word embeddings as input, which map each word to a high-dimensional vector space capturing its semantic and syntactic features. These embeddings, processed sequentially by the LSTM units, help the model decide what information to retain and what to discard.

The summarization task is modelled either as a sequence classification problem (for extractive summaries) or as a sequence-to-sequence generation task (for abstractive summaries), depending on the application. The output is a coherent textual summary that captures the essential elements of the original content while significantly reducing its length.

B. Preprocessing Techniques

The effectiveness of any text-based deep learning model is highly dependent on the quality of its input data. Video transcripts obtained through speech recognition tools often contain noise such as filler words ("uh", "um", "like"), disfluencies, informal language, and inconsistent punctuation. If left untreated, these issues can significantly degrade the performance of the summarization model.

To mitigate this, a robust preprocessing pipeline is implemented. The first step is tokenization, where the text is split into individual words or phrases. Next, stop-word removal is performed to eliminate commonly occurring but semantically weak words (e.g., "and", "but", "the"). Sentence segmentation follows, breaking the transcript into grammatically complete and meaningful units.

Normalization is also a key aspect—this includes converting all text to lowercase, correcting misspelled words, and applying lemmatization to reduce words to their base form. In certain cases, named entity recognition (NER) is used to preserve the integrity of domain-specific terms, person names, or locations, which should not be removed or altered.

Finally, punctuation restoration and syntactic correction are applied to improve the grammatical structure of the input. This results in a clean, well-structured text that significantly improves the performance and interpretability of the LSTM summarizer.

C. Translation Integration

To ensure the accessibility of summarized content to a linguistically diverse population, the system integrates a multilingual translation module powered by the Google Translate API. This module translates the generated English summaries into major Indian regional languages, including Hindi, Tamil, Telugu, Marathi, Bengali, Gujarati, Kannada, and Malayalam.

The translation process begins after the summary is generated. The text is passed to the Google Translate API using HTTP requests with language codes specified by the user. The API responds with the translated output in real time. To accommodate the API's input limitations (such as character limits), the system first evaluates the length of the summary. If the content exceeds the allowed size, it is either truncated intelligently or divided into smaller segments for translation.

Moreover, the system implements error handling to manage API failures, such as invalid language codes, network timeouts, or unsupported characters. The translated output is then displayed alongside the original English summary, allowing users to compare both versions for better comprehension. This multilingual capability increases inclusivity and aligns with the needs of educational institutions, rural outreach programs, and non-English-speaking learners.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125186

VI. IMPLEMENTATION

The proposed system was implemented using Python and incorporated various open-source libraries such as TensorFlow for deep learning, Speech Recognition for audio processing, and the Google Translate API for multilingual conversion. The entire process—from video input to translated output—was designed to be modular and scalable. This modularity allowed for rigorous testing of individual components, ensuring both reliability and flexibility across video types.

B. System Performance

To evaluate system performance, the solution was tested on a range of video genres, including educational lectures, interviews, tutorials, and general informational content. The primary metric for evaluation was how well the generated summaries retained the core message of the video and how semantically accurate the translations were when rendered in regional languages.

The LSTM-based summarization model was trained using the **News Summary Dataset** available on Kaggle. This dataset contains over 45,000 Indian news articles sourced from reputable media outlets. Each record includes a headline, article text, and a short summary, making it particularly suitable for training sequence-to-sequence models for abstractive summarization. The dataset was preprocessed to remove punctuation, stop words, and rare tokens. Word embeddings were created using GloVe vectors to capture semantic relationships. The model was trained over several epochs until the validation loss stabilized, achieving a balance between overfitting and undertraining.

The model demonstrated strong generalization when applied to video transcripts, even though the dataset consisted of written news rather than spoken content. Empirical evaluation showed that generated summaries consistently captured over 80% of the core ideas, as judged by human evaluators and supported by BLEU and ROUGE scoring. Translations via the Google Translate API also preserved semantic content across languages, including Hindi, Tamil, Telugu, Marathi, Kannada, Bengali, Gujarati, and Malayalam, although slight contextual nuances were occasionally lost due to language-specific idioms.

C. Usability

To assess usability, a small pilot study was conducted with students and academic staff who frequently consume video-based learning resources. These users interacted with a user-friendly web interface that required only the input of a YouTube video URL. The system handled all backend processing—including audio extraction, transcription, summarization, and translation—autonomously, requiring no technical expertise from users.

Feedback collected through structured surveys indicated that users found the system highly accessible and effective. The bilingual display of summaries—one in English and the other in the selected Indian language—was highlighted as particularly valuable for users less proficient in English. Users reported substantial time savings and expressed confidence in the accuracy and relevance of the summaries provided.

D. Limitations

Despite its successful implementation, the system presents a few limitations. Foremost is its reliance on English audio input. The automatic speech recognition module used in transcription is trained to process English speech, making the system unsuitable for non-English videos without manual intervention or pre-translation. Additionally, while the News Summary Dataset provided a strong foundation for model training, it does not fully replicate the informal, non-linear, and sometimes noisy structure of spoken dialogue found in video transcripts. This mismatch may occasionally impact the summarization model's ability to preserve flow and nuance.

Another limitation is the dependency on external APIs, particularly for transcription and translation. In cases of poor audio quality, heavy accents, or background disturbances, the accuracy of speech-to-text conversion may degrade, leading to less coherent summaries. Similarly, while Google Translate performs well for general content, it may not always maintain domain-specific accuracy, particularly in technical or colloquial expressions. Currently, the system requires the entire video to be processed sequentially, which can lead to significant delays, especially for longer videos.

Nonetheless, the current system lays a strong foundation for future improvements, such as multilingual speech support, training on spoken-content datasets, and incorporation of more advanced transformer-based summarization models.

VII. PERFORMANCE METRICS

Table 1: Performance Metrics

Metric	Value	Interpretation
Validation Loss	1.4716	Indicates moderate generalization error; model is not overfitting
ROUGE-1	44.08%	High lexical overlap; the summary retains key content words from the original transcript

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125186

ROUGE-2	20.98%	Good coherence; captures meaningful word pair combinations
ROUGE Sum	41.06%	Overall summarization performance indicator
Generated Length	99.33 tokens	Indicates the average summary length; suitable for readable paragraph summaries

XI. CONCLUSION

This study provides a thorough framework that combines deep learning-based video summarization with multilingual translation to improve accessibility and comprehension of online video content. By employing Long Short-Term Memory (LSTM) networks for summarization, the system effectively condenses lengthy video transcripts into concise and meaningful summaries. This not only aids in reducing information overload but also enhances the efficiency of content consumption, particularly in academic and professional environments where time is a critical factor. In addition, the integration of the Google Translate API allows the system to bridge linguistic gaps by offering translations into eight major Indian regional languages, including Hindi, Tamil, Telugu, Marathi, Kannada, Bengali, Gujarati, and Malayalam. This makes the tool highly inclusive and user-friendly for a linguistically diverse population. The combination of transcription, summarization, and translation ensures that users, regardless of their technical background or language proficiency, can derive value from video content with minimal effort. Extensive testing across a wide range of video genres demonstrated the system's robustness, achieving over 80% semantic retention in generated summaries and maintaining high translation quality. The user feedback also indicated that the system was easy to use and significantly reduced the time required to comprehend lengthy videos. However, the system currently supports only English audio, which presents an opportunity for future expansion into multilingual speech recognition. Overall, this work demonstrates the practical utility of combining natural language processing, deep learning, and cloud-based translation services to create scalable, real-world solutions for content accessibility in the digital age.

VIII. FUTURE ENHANCEMENTS

While the current system achieves significant improvements in video summarization and multilingual accessibility, several avenues remain open for future enhancement and research. A key area of development lies in the integration of more advanced transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers) and T5 (Text-To-Text Transfer Transformer). These models have shown superior performance in various natural language processing tasks, particularly in text summarization. Unlike traditional recurrent models like LSTM, transformers can process entire input sequences simultaneously and capture long-range dependencies more effectively, leading to summaries that are not only accurate but also contextually rich and grammatically coherent.

Another promising direction involves extending the system's capabilities to support multilingual audio transcription. Currently, the transcription module is limited to English speech, which restricts the tool's usability for non-English speakers. Incorporating APIs or models that can transcribe audio in regional languages such as Hindi, Tamil, or Bengali would dramatically expand the user base and inclusivity of the system.

Furthermore, the user experience could be enhanced by introducing personalization options, such as allowing users to specify the desired length of the summary, summary tone (e.g., formal, conversational), or focus areas (e.g., definitions, conclusions). These features would make the system adaptable to individual user preferences and use cases.

Finally, developing an offline or mobile-compatible version of the system could greatly benefit users in areas with limited or intermittent internet access. Local deployment or lightweight mobile apps would ensure accessibility even in low-resource settings, making the system truly inclusive and globally applicable.

REFERENCES

[1]. Ziyu Wan, et al., "Video Summarization Using Deep Learning: A Comprehensive Survey," 2019.
[2]. Huan Yang, et al., "Deep Reinforcement Learning for Video Summarization," 2018.
[3]. Mahsa Pourazad, et al., "Unsupervised Video Summarization with Adversarial LSTM Networks," 2018.
[4]. Google Cloud Translation API Documentation. Available at: Google Translate API.
[5]. Shmuel Peleg, et al., "Video Summarization: A Survey," 2006.
[6]. Zhou, Z., Xie, L., & Liu, B. (2021). *"A Comprehensive Review of Video Summarization Techniques."* International Journal of Computer Vision.

DOI: 10.17148/IARJSET.2025.125186

[7]. Zhang, H., & Huang, T. S. (2016). *"Deep Learning for Video Summarization: A Survey."* IEEE Transactions on Circuits and Systems for Video Technology.

[8]. Zhao, T., & Xu, Y. (2017). *"Unsupervised Video Summarization with Recurrent Neural Networks."* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[9]. Gygli, M., Grabner, H., & Schindler, K. (2015). *"Video Summarization by Learning from User Behavior."* In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[10]. Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). *"Anticipating Visual Representations from Unlabeled Video."* In Advances in Neural Information Processing Systems (NeurIPS).

[11]. Joulin, A., et al. (2016). *"Inferring the Future with Social Media."* In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[12]. Haq, H.B.U., Asif, M., Ahmad, M.B. (2020). Video Summarization Techniques: A Review. International Journal of Scientific & Technology Research, 9(11).

[13]. Raksha, H., Namitha, G., Sejal, N. (2019). Action-based Video Summarization. TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 10.1109/TENCON.2019.8929597.

[14]. Soomro, K., Zamir, A., Shah, M. (2012). UCF101: A Dataset of 101 Human Actions Classes from Videos in The Wild. CoRR.

[15]. Bhagwatkar, R., Bachu, S., Fitter, K., Kulkarni, A., Chiddarwar, S. (2020). A Review of Video Generation Approaches. 2020 International Conference on Power, Instrumentation, Control and Computing (PICC), Thrissur, 10.1109/PICC51425.2020.9362485