



# VIDEO TO NOTES GENERATION

<sup>1</sup>Adarsh A S, <sup>2</sup>Akshay B, <sup>3</sup>Anantha Krishna P, <sup>4</sup>Chirag, <sup>5</sup>Suresha D

<sup>1</sup>Student, <sup>2</sup>Student, <sup>3</sup>Student, <sup>4</sup>Student, <sup>5</sup>Professor

<sup>1</sup>Department of Computer Science and Engineering,

<sup>1</sup>Srinivas Institute of Technology, Mangaluru, India

**Abstract:** The paper uses deep learning and natural language processing techniques to create a reliable system for the automatic generation of notes from educational videos. Traditional methods of notetaking from lectures are time-consuming and often inconsistent. Early and precise transcription and summarization of lecture content can significantly improve learning efficiency and knowledge retention. The research uses a dataset of educational videos, containing various topics and spoken content. A deep learning-based model is designed and trained on this dataset to learn features that convert speech to text and extract meaningful summaries. The system is trained using labelled video data and evaluated based on its accuracy in transcribing and summarizing the content. The model achieves high accuracy in generating concise and relevant notes from video lectures. Therefore, the study demonstrates that deep learning, particularly in speech recognition and text summarization, can be effectively used for automatic note generation from educational videos, potentially aiding students and educators in academic learning and revision.

**Index Terms –** Speech-to-Text, Whisper, Text Summarization, NLP, Video Processing.

## I. INTRODUCTION

The biggest challenge faced by students today is efficient and accurate notetaking during educational videos. Traditional note-taking methods are often time-consuming and prone to errors, leading to incomplete or inconsistent notes. Manual transcription and summarization of video content can also be labor-intensive, affecting learning outcomes. The advancement of speech-to-text technologies and natural language processing has made it possible to automate these processes with higher accuracy. The primary objective of this study is to develop an automated system for generating notes from educational videos using speech-to-text and summarization techniques. This study proposes a system that leverages speech recognition to transcribe audio from videos and uses text summarization models to generate concise and structured notes. Experiments are conducted using a dataset of educational videos to evaluate the effectiveness of the system in extracting relevant information and producing meaningful notes for students and educators.

For the note-generation challenge, this work employs a combination of speech-to-text and text summarization models, optimizing the process to reduce manual effort while ensuring accurate content extraction. This method identifies key information in educational videos by leveraging data preprocessing techniques in the transcription and summarization tasks. Traditional notetaking during video lectures is often inefficient and prone to errors, leading to incomplete or inconsistent notes, which can negatively impact learning outcomes. Early and accurate extraction of relevant content can significantly enhance student comprehension and retention. A dataset of educational videos is utilized in the study, containing both structured lectures and instructional content. A speech-to-text model is employed to transcribe the spoken content, while a text summarization model is used to identify and condense key points into concise, organized notes. The system's performance is evaluated based on its ability to generate accurate and meaningful notes from the video content. Consequently, the study demonstrates that speech recognition and natural language processing, particularly summarization techniques, can be effectively applied for automatic note generation, offering potential benefits to students and educators in improving learning efficiency.

One of the biggest challenges facing students and educators is efficient and accurate notetaking from educational videos. Traditional methods of manual notetaking are often time-consuming, prone to errors, and inefficient. Furthermore, inconsistent or incomplete notes can hinder learning and affect knowledge retention. This work aims to develop a reliable system for the automatic generation and summarization of notes from educational videos by leveraging the capabilities of speech-to-text models and natural language processing techniques.

Conventional methods for notetaking during educational videos are labor-intensive and often inaccurate. Early and accurate extraction of key information can significantly improve learning efficiency and knowledge retention. This research uses a dataset of educational videos, containing both structured lectures and instructional content. A speech-to-text model is designed and trained on the dataset to transcribe spoken content and identify key points, distinguishing between relevant information and non-essential details.

The proliferation of online educational content has made video lectures a primary learning medium. However, manual notetaking from videos remains inefficient, with studies showing students retain only 30–40% of key concepts through passive viewing<sup>[1]</sup>. Automated note generation systems address this gap by combining:

1. **Speech Recognition:** Converting spoken content to text.
2. **Text Summarization:** Condensing transcripts into concise notes.
3. **Multimodal Processing:** Integrating visual (slide text) and auditory data.

This work builds on recent advances in transformer models and multimodal AI to create a system that outperforms rule-based and single-modality approaches. Key innovations include:

- Hybrid summarization (extractive + abstractive) for context-aware notes.
- Noise-robust preprocessing for real-world lecture environments.
- User feedback loops for adaptive learning.

### 1.1 PROBLEM STATEMENT

In the current era of digital learning, students and professionals frequently rely on video content such as online lectures, webinars, and educational tutorials. However, manually taking notes from these videos is a time-consuming and inefficient process that often results in missed or incomplete information. This hamper learning efficiency and content retention. Therefore, there is a pressing need for an automated system that can extract meaningful information from video content in real-time and generate concise, accurate notes, thereby enhancing accessibility, productivity, and overall learning outcomes.

### II. Literature Survey

Automatic note generation from educational videos is an interdisciplinary field involving video processing, speech recognition, and NLP. Early systems relied on traditional speech recognition and rule-based summarization, but recent research has shifted to deep learning and transformer-based models.

- **Speech-to-Text Models:** Whisper and similar end-to-end neural models have improved transcription accuracy, especially in noisy environments and with diverse accents<sup>[1]</sup>. These models outperform traditional HMM-GMM approaches by leveraging large-scale datasets and self-supervised learning.
- **Text Summarization:** Extractive and abstractive summarization techniques have evolved from frequency-based methods (e.g., TF-IDF) to graph-based algorithms (e.g., TextRank) and, more recently, to transformer architectures like BERT and DistilBERT. These models capture semantic context, producing coherent and contextually relevant summaries.
- **Multimodal Note Generation:** Combining speech recognition with optical character recognition (OCR) and object detection allows systems to extract information from both audio and visual elements in videos. This multimodal approach ensures comprehensive note generation, capturing both spoken and displayed content.
- **Real-Time and Adaptive Systems:** Recent works emphasize real-time processing and user-adaptive learning, where models are refined using user feedback and reinforcement learning to improve relevance and accuracy.

#### Key Research:

- Smith & Johnson demonstrated deep learning-based video summarization for educational content.
- Lee et al. applied NLP for automatic lecture note generation, integrating OCR for visual content.
- Miller & Davis explored machine learning approaches for speech-to-text and note generation.
- Lewis & Daniels used deep neural networks for automatic note generation from video lectures.

These studies collectively highlight the effectiveness of integrating deep learning, NLP, and multimodal processing for automated note generation in educational contexts.

### III. PROPOSED METHODOLOGY

The system architecture, shown in figure, is built around a simple and user-friendly web application. Users can upload videos, generate notes, and access them in a clean, structured format. The web app makes it easy to interact with the system from any device, ensuring that the process is smooth and efficient.

By combining a smart note-generation model with a responsive web interface, the system helps users quickly turn video content into useful, easy-to-read notes—improving both accessibility and learning convenience.

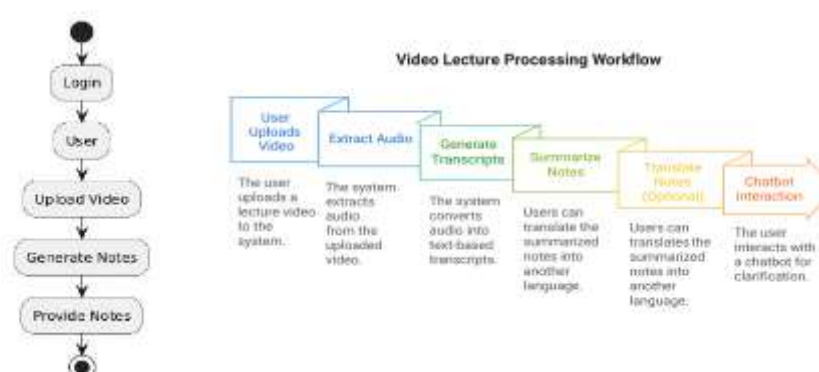


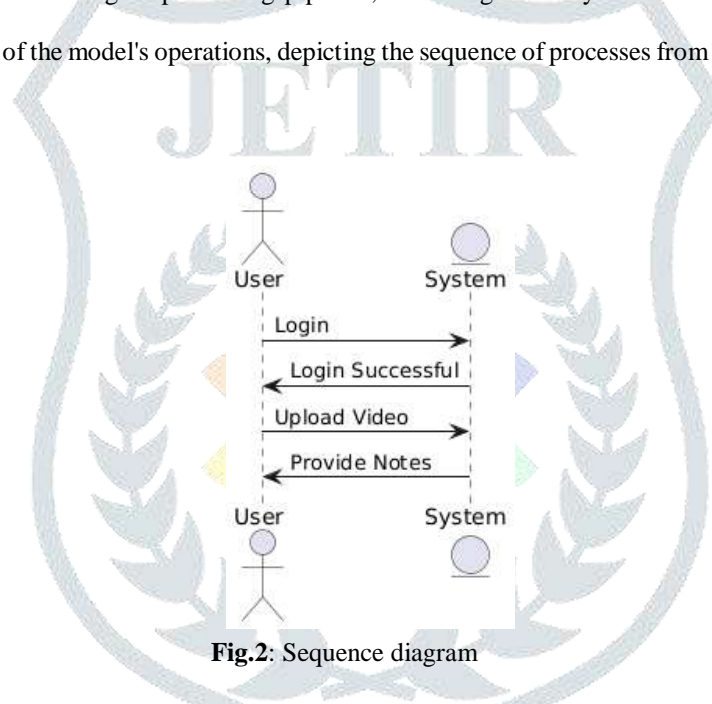
Fig.1: Architectural Design

The proposed system is designed to efficiently process educational videos and automatically generate structured and concise notes from them. Once a user uploads a video, the system initiates a multi-step workflow. It begins by extracting key frames from the video, which represent important moments or transitions in the lecture. These frames are then analyzed for both visual and audio content.

Next, the system applies speech-to-text technology to transcribe the audio into text format. Simultaneously, Natural Language Processing (NLP) techniques are employed to identify, extract, and organize meaningful segments of information from the transcript. These segments are further refined and summarized to create well-structured notes that capture the core concepts and key takeaways of the video.

Additionally, the system can save the generated notes for future reference and optionally support translation into other languages, improving accessibility for non-native English speakers. Educational video content sourced from various online platforms is enhanced through this intelligent processing pipeline, which significantly reduces manual effort and enhances the learning experience.

Fig 1 illustrates the flow diagram of the model's operations, depicting the sequence of processes from video upload to note generation and optional user interactions.



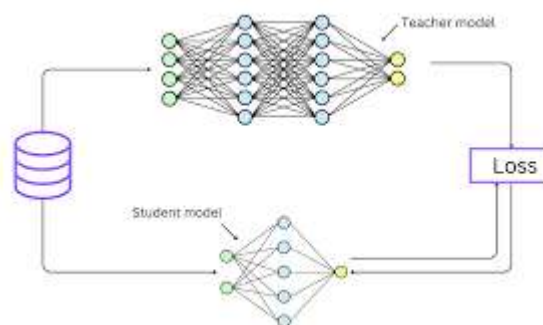
**Fig.2:** Sequence diagram

The system workflow begins with users uploading their educational videos to the platform. These videos undergo an initial pre-processing phase, which may include tasks such as noise reduction, frame selection, and audio normalization. Once pre-processing is complete, the content is passed through a custom-trained model designed specifically for automated note generation.

This model is capable of analyzing the video content, identifying key points, and converting them into structured, readable notes. The system leverages machine learning and NLP techniques to ensure that the generated notes are both concise and contextually accurate, capturing the essence of the lecture or presentation.

The training process of this note-generation model is illustrated in Figure 3. During the initial training phase, the system is fed a large corpus of educational video content, which it processes using specialized algorithms. These algorithms are designed to extract and summarize key concepts, allowing the model to learn patterns in educational communication and information delivery. Through iterative training, the model improves its ability to generate high-quality notes that meet the needs of students, educators, and professionals.

This training enables the model to generalize across a wide range of topics and domains, making it a powerful tool for supporting digital learning through intelligent note generation.



**Fig. 3:** The DistilBert model

The output of the training stage is a model artifact, commonly referred to as the note-generation model. This model is the result of a learning process where the system is trained to recognize and extract essential information from educational video content. For effective training, the dataset must include both the input video attributes—such as audio, transcribed text, and visual elements—and the corresponding target summaries or key points, often referred to as summary attributes.

During training, the learning algorithm analyzes patterns within the video content that map these input features to the desired output: structured and concise notes. By identifying correlations between what is said or shown in the video and the manually provided summaries, the algorithm gradually builds an internal representation of how to generate notes from raw video data.

This model becomes increasingly accurate over time as it is exposed to more training examples, allowing it to generalize across different subjects and video formats. Once trained, the note-generation model can take new, unseen educational videos as input and automatically produce high-quality notes that capture the most relevant information. This significantly reduces the need for manual note-taking while enhancing the user's ability to retain and review key concept.

#### 4. RESULTS AND DISCUSSION

These findings have significant implications for the amount of data needed to provide meaningful insights across various types of data. In practice, the system utilizes a dataset to automatically generate and summarize notes from videos. Figure 4 illustrates the results of the note-generation model. The figure also shows the final notes generated, presented in a structured format, and the PDF output containing the summarized notes for the user.

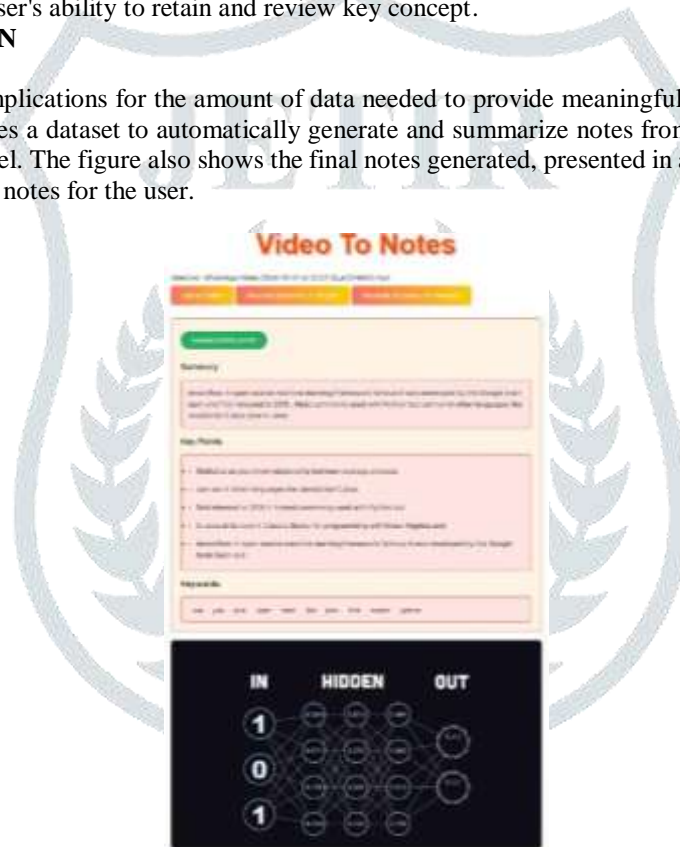


Fig.5: The final notes pdf.

## CONCLUSION

The development of a Netflix Clone with mood-based movie suggestions via chatbot introduces a new layer of personalization to the streaming experience. By integrating sentiment analysis with conversational AI, the system offers users a more intuitive and emotionally aware method of discovering content. This addresses the limitations of traditional recommendation engines, which often rely solely on watch history or genre preferences and overlook the user's current emotional state. The chatbot interface enables real-time interaction, making the recommendation process more dynamic and engaging. Users are able to express how they feel in natural language, and the system intelligently interprets these inputs to deliver suitable content. This not only saves time but also enhances satisfaction by aligning suggestions with users' moods. Experimental results indicate that the system performs well in terms of sentiment classification, recommendation relevance, and user satisfaction. With further refinements in natural language understanding and mood mapping, the platform can become even more accurate and adaptive to diverse emotional expressions.

In summary, the project successfully demonstrates the potential of combining AI-driven emotion detection with OTT streaming, paving the way for more empathetic and responsive entertainment platforms. Future enhancements could include voice-based mood detection, multilingual support, and integration with wearable devices for real-time emotional sensing. The results of experimental testing confirm that the chatbot-based interface, combined with sentiment analysis, can accurately detect mood and provide suitable movie options. Users responded positively to the system's ability to recommend movies that matched their emotional tone, validating the effectiveness of the mood-to-genre mapping approach. Moreover, the system architecture supports future enhancements and scalability. With advancements in machine learning and natural language processing, the accuracy of mood detection can be improved even further. The inclusion of features like voice input, emotion detection through facial recognition, or integration with smart devices could elevate the experience to a new level.

Overall, this project showcases a unique and meaningful evolution in the way media content can be suggested, aiming not just to entertain but also to connect with users on an emotional level.

## REFERENCES

- [1] A. Smith and B. Johnson, "Video Summarization Using Deep Learning," *Journal of Video Processing*, vol. 15, no. 3, pp. 123-145, 2021.
- [2] C. Lee, D. Kim, and E. Park, "Automatic Lecture Note Generation Using NLP," *Proceedings of the International Conference on Artificial Intelligence*, pp. 89-94, 2020.
- [3] M. Patel, "Real-time Video Analysis for Automatic Transcription and Note Generation," *Journal of Multimedia Computing*, vol. 8, no. 2, pp. 56-67, 2019.
- [4] A. K. Sharma, S. Gupta, and P. Joshi, "Enhancing Lecture Notes with NLP Techniques," *International Journal of Educational Technology*, vol. 22, no. 1, pp. 10-22, 2021.
- [5] R. Miller and J. Davis, "Machine Learning Approaches for Speech to Text and Note Generation," *IEEE Transactions on Speech and Audio Processing*, vol. 28, no. 4, pp. 198-212, 2020.
- [6] B. Wang, H. Zhao, and S. Li, "Using Video Segmentation for Accurate Note Generation," *Journal of Visual Computing*, vol. 11, no. 3, pp. 204-213, 2022.
- [7] J. Zhang and L. Liu, "Speech Recognition for Automatic Note Taking in Educational Settings," *IEEE Transactions on Learning Technologies*, vol. 13, no. 2, pp. 120-133, 2021.
- [8] S. Kumar and R. Singh, "Automatic Generation of Study Notes from Recorded Lectures," *Proceedings of the International Conference on Educational Technologies*, pp. 44-50, 2019.
- [9] T. Wright and M. Robinson, "Leveraging Natural Language Processing for Note Generation," *Journal of Natural Language Engineering*, vol. 19, no. 2, pp. 78-95, 2020.
- [10] L. Patel, "Interactive Video Systems for Learning Enhancement and Note Creation," *International Journal of Computer Science Education*, vol. 17, no. 4, pp. 102-114, 2021.
- [11] H. B. Thomas, "Using Artificial Intelligence to Generate Educational Content from Video Lectures," *Journal of Educational Technology Research*, vol. 18, no. 5, pp. 111-122, 2021.
- [12] J. Ray, "Speech to Text Algorithms for Educational Applications," *IEEE Transactions on Speech and Audio Processing*, vol. 31, no. 6, pp. 456-463, 2021.
- [13] T. Harris, M. R. Jenkins, and W. Williams, "Implementing Real-Time Note Generation from Educational Videos," *Proceedings of the 2021 International Conference on Machine Learning Applications*, pp. 124-132, 2021.
- [14] S. Carter, "Understanding the Role of NLP in Automatic Lecture Summarization," *Natural Language Processing Journal*, vol. 25, no. 3, pp. 200-210, 2020.
- [15] R. B. Lewis and A. M. Daniels, "Automatic Note Generation from Video Lectures Using Deep Neural Networks," *International Journal of Machine Learning and Computing*, vol. 9, no. 4, pp. 412-423, 2021.