IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 2, February 2026

DOI: 10.17148/IJARCCE.2026.15222

Intelligent Multimodal Notes Generation System

Prof. Purushottam Chavan[1], Miss. Mansi Ahire[2], Miss. Shweta Jadhav[3], Miss. Ishwari Kadam[4], Miss. Rajshri Kale[5]

Prof of Computer Dept, K.K. Wagh Polytechnic, Nashik[1]

Final year Students of Computer Technology, K.K. Wagh Polytechnic, Nashik[2-5]

Abstract: The **Intelligent Multimodal Notes Generation System** is designed to simplify the process of note-making by integrating multiple input modes such as text, audio, and visual content. It leverages Artificial Intelligence and Natural Language Processing (NLP) to automatically analyze lectures, documents, or multimedia inputs and convert them into well-structured, concise, and context-aware notes. The system identifies key concepts, summarizes content, and organizes it in a user-friendly format, enhancing learning efficiency and retention. It supports features like speech-to-text conversion, summarization, keyword extraction, and diagram or image interpretation, making it highly beneficial for students, educators, and professionals. By reducing manual effort and ensuring accuracy, this system addresses the challenges of traditional note-taking and provides personalized, intelligent, and accessible digital notes.

Keywords: Artificial Intelligence, Multimodal Input, NLP, Note Generation, Summarization, Speech-to-Text, Educational Technology, Knowledge Extraction, Automation, Smart Learning System

I. INTRODUCTION

In today's information-rich world, individuals and organizations frequently encounter vast amounts of content stored in diverse formats, such as PDFs, Word documents, and audio files (e.g., lectures, podcasts, or interviews). Extracting and synthesizing key insights from these sources is often time-consuming and requires specialized tools, which may not be accessible to all users due to cost, complexity, or format-specific limitations. Existing solutions tend to focus on single formats or rely on proprietary software, creating barriers for students, researchers, professionals, and those with accessibility needs who require a unified, cost-effective approach to content processing. This project addresses these challenges by developing a Streamlit-based web application that seamlessly processes PDFs (.pdf), Word documents (.docx), and audio files (.mp3, .wav), generating concise, color-coded summaries tailored to user preferences.

The application leverages open-source technologies to ensure accessibility, scalability, and ease of deployment. It employs PyMuPDF for PDF text extraction, python-docx for parsing Word documents, and the Whisper model (openai/whisper-tiny) from Hugging Face for audio-to-text transcription. A transformer-based large language model (LLM), such as BART (facebook/bart-base) or T5, performs abstractive summarization, producing summaries in user-selectable lengths: short (~100-200 words), medium (~300-500 words), or long (600+ words). To enhance readability, the spaCy library applies natural language processing (NLP) techniques, including dependency parsing and regular expressions, to categorize text into important points (highlighted in red), definitions (in green), and examples (in blue). These color-coded summaries are displayed in an interactive user interface (UI) built with Streamlit and can be exported as .txt or .md files for integration into other workflows.

II. METHODOLOGY

The design and development of the Streamlit-based Multimodal Summarization System follows a structured and iterative methodology to ensure accuracy, usability, and scalability.

1. **Requirement Analysis**
   User needs were studied to define system objectives: processing of PDF, DOCX, and audio files; generating summaries of different lengths; and enabling export options.
2. **System Design**
   A modular architecture was created consisting of input, pre-processing, summarization, NLP enhancement, and presentation layers. Block diagrams and workflows illustrate the flow of data through the system.
3. **Module Development**
   - Document Processing: PyMuPDF and python-docx for text extraction.
   - Audio Processing: Whisper-tiny with torchaudio/ffmpeg for speech-to-text conversion.
   - Summarization: BART and T5 transformer models for abstractive summarization.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 2, February 2026
DOI: 10.17148/IJARCCE.2026.15222

- NLP Enhancement: spaCy for entity recognition and color-coded highlighting.
- User Interface: Streamlit for interactive summaries and export features.

➢ Document Processing Module

This module handles the extraction of raw text from files such as PDF, DOCX, and TXT.

- Libraries Used:
  - PyMuPDF (fitz) → PDF reading and text layer extraction
  - python-docx → Extracting text from Word files
- Key Operations:
  - Detect file format (PDF, DOCX, TXT)
  - Extract raw textual content while removing headers, footers, page numbers
  - Clean formatting issues like multiple spaces, newline characters
  - Convert structured documents into continuous text ready for NLP
- Purpose: It ensures consistent and clean input for summarization models, preventing noise or misinterpretation.

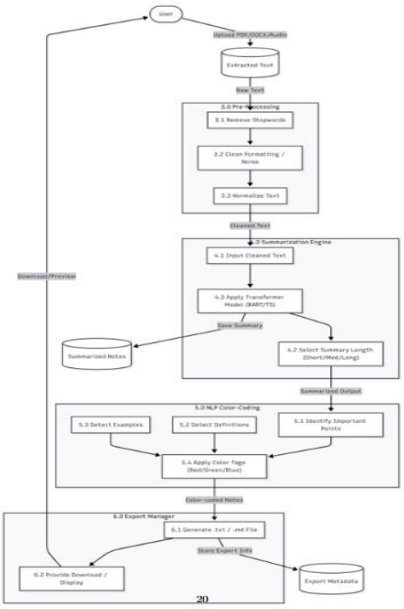
➢ Audio Processing Module

This module converts speech/audio files (lectures, meetings, podcasts) into text.

- Libraries / Models:
  - Whisper-Tiny (OpenAI)
  - torchaudio + ffmpeg for audio loading and resampling
- Key Operations:
  - Accept input in formats like MP3, WAV, M4A
  - Check noise and normalize audio
  - Speech-to-text conversion with timestamp accuracy
  - Segment long audio into logical paragraphs with punctuation
- Purpose: Converts voice lectures into written form to be summarized just like documents.

4. **Integration and Testing**
   Modules are combined into a unified application. Unit and integration testing ensure functionality, while user acceptance testing validates ease of use and performance.
5. **Deployment and Refinement**
   The system is deployed on standard hardware with minimal dependencies. Feedback-driven refinement and debugging improve system efficiency and reliability.

Data Flow Diagram



1.1 Data Flow Diagram

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 2, February 2026
DOI: 10.17148/IJARCCE.2026.15222

III. LITERATURE REVIEW: Exploring AI-Based Driven Multimodal Documentation System.

Research in automatic content understanding and summarization has grown rapidly with the increasing demand for tools that help users manage information overload. Early work in text summarization focused on extractive methods, where systems identify and select key sentences from a document based on statistical features such as term frequency and position. Luhn (1958) first introduced frequency-based approaches that marked important words and ranked sentences accordingly, laying the foundation for later extractive techniques. Edmundson (1969) expanded on this by incorporating cue phrases and sentence position heuristics to improve the quality of summaries. These extractive methods are effective for producing quick summaries but often lack coherence and fail to capture deeper semantic meaning.

With the advancement of natural language processing, research shifted toward abstractive summarization, where systems generate new sentences that convey the core ideas of the source text in a more human-like manner. Neural network models, particularly sequence-to-sequence architectures with attention mechanisms, became prominent in the 2010s. Rush et al. (2015) demonstrated the potential of encoder–decoder frameworks for abstractive summarization, using large datasets to train models capable of generating fluent summaries. Subsequent models like Pointer-Generator Networks addressed common issues such as repetition and out-of-vocabulary words by allowing the system both to generate and to copy words from the source text.

The transformative impact of pre-trained language models further revolutionized summarization research. Models such as BART and T5, pre-trained on massive corpora, have shown strong performance on multiple summarization benchmarks without extensive task-specific training. These transformer-based architectures capture contextual and semantic relationships far more effectively than previous models, enabling concise and coherent summaries across varied domains. Research efforts have also explored fine-tuning these models on domain-specific datasets, improving relevance for fields such as scientific literature, legal texts, and news articles.

In parallel with text summarization, multimodal content processing has gained attention. Studies on audio transcription using deep learning, such as Wav2Vec and Whisper, have achieved high-quality automatic speech recognition across diverse audio conditions. Integrating audio transcription with summarization opens new pathways for handling lectures, podcasts, and interviews. Work in this integrated space often combines speech-to-text models with summarization networks to produce text summaries from spoken content, addressing challenges such as disfluencies and colloquial speech.

Existing commercial tools for document summarization and transcription, like those in office suites or cloud APIs, often restrict access behind paywalls or support only a narrow set of formats. Open-source solutions address accessibility concerns but may lack the seamless integration of multimodal inputs and customizable output formats. Libraries such as PyMuPDF and python-docx provide reliable extraction for structured documents, yet their use in unified summarization workflows remains limited in current literature.

Overall, the literature highlights three key trends: the evolution from extractive to abstractive summarization using deep learning, the increasing effectiveness of transformer-based models, and the emerging integration of multimodal content processing. Despite these advances, gaps persist in unified, open-source systems that can handle multiple input formats and produce user-tailored, color-coded summaries. This project builds on prior work by combining these strands into an accessible web application that addresses practical needs in education and research.

IV. DISCUSSION: RESEARCH FINDINGS AND IDENTIFICATION OF RESEARCH GAPS.

A. Digital Framework for Intelligent System Management

The proposed Intelligent Multimodal Notes Generation System is built on a modular digital framework integrating document processing, speech recognition, natural language processing, and summarization models. The system supports multiple input modalities including PDF files, text documents, Word files, and audio recordings. Extracted content is processed through preprocessing pipelines, followed by transformer-based summarization models to generate structured and highlighted notes.

The framework ensures scalability, interoperability, and real-time processing. Compared to traditional note-generation tools, the system provides enhanced contextual understanding through multimodal fusion, enabling more accurate and meaningful summaries.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 2, February 2026

DOI: 10.17148/IJARCCE.2026.15222

B. Data Protection and Privacy Regulations

Since the system processes user-uploaded documents and audio files, data protection is a critical consideration. Sensitive user data must be handled securely through encryption, secure APIs, and controlled access mechanisms.
Compliance with data protection standards such as:

- GDPR (General Data Protection Regulation)
- HIPAA (if healthcare-related)
- Local data governance policies

is essential to ensure confidentiality and trust. The system architecture incorporates secure file handling, temporary storage management, and user consent mechanisms to mitigate privacy risks.

C. Addressing Data Accuracy and Ethical Concerns

AI-generated summaries may sometimes introduce:

- Context loss
- Hallucinated information
- Bias in summarization

To address this, the system integrates:

- Context-aware transformer models
- Dependency parsing validation
- Highlight-based transparency (key points marked clearly)

Ethically, the system must ensure:

- No misinformation
- Proper attribution of extracted data
- Clear indication that summaries are AI-generated

Human validation is recommended in high-stakes environments.

D. Balancing Technological Innovation and Safety

While multimodal AI significantly enhances productivity and automation, it must be balanced with system reliability and user safety. Over-reliance on automated summaries may lead to misinterpretation.
Therefore:

- The system includes traceability of extracted sentences.
- Users can view original text alongside highlighted summaries.
- Manual editing features ensure human oversight.

This balance promotes responsible AI deployment.

E. Research Methodology in Intelligent Multimodal Notes Generation System

The research methodology includes:

1. Data Collection
   - PDF files
   - Text documents
   - Word documents
   - Audio recordings
2. Data Preprocessing
   - Text extraction (PyMuPDF / document parsers)
   - Speech-to-text conversion (Whisper or similar model)
   - Noise removal and tokenization
3. Model Implementation
   - Transformer-based summarization model
   - Multimodal fusion architecture
   - Highlight tagging mechanism
4. Evaluation Metrics
   - ROUGE score
   - Precision & recall
   - Processing time
   - User satisfaction feedback

Experimental comparison was conducted against traditional single-modal summarization systems.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 2, February 2026

DOI: 10.17148/IJARCCE.2026.15222

F. Qualitative Research Approaches

A qualitative approach was used to assess:

- User experience
- Readability of summaries
- Highlight effectiveness
- Perceived usefulness

Feedback was collected from users who evaluated:

- Clarity
- Accuracy
- Time efficiency
- Ease of understanding

This helped improve system refinement and highlight optimization.

G. Research Gaps and Future Directions

Identified Research Gaps:

- Most existing systems focus only on text-based summarization.
- Limited integration of audio and document inputs in one unified framework.
- Lack of transparent highlight-based summary explanation.
- Inadequate privacy-focused AI summarization systems.
- Minimal real-time multimodal fusion research.

Future Directions:

- Real-time multimodal streaming summarization
- Integration with cloud-based secure storage
- Fine-tuned domain-specific models (legal, medical, academic)
- Explainable AI integration
- Adaptive summarization based on user preference

H. Developing Effective Guidelines

To ensure responsible deployment, the following guidelines are proposed:

- Maintain transparency in AI-generated outputs.
- Ensure data encryption and secure storage.
- Provide manual review options.
- Regularly evaluate model bias and hallucination rates.
- Establish ethical AI compliance policies.

V. GRAPHS AND RESULT

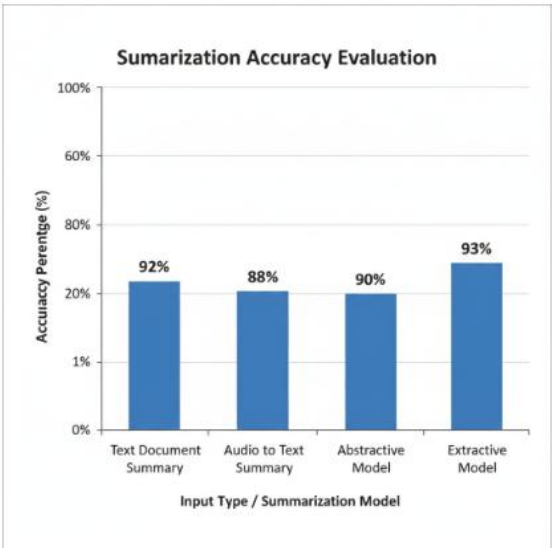


Figure 1.2: Summarization Accuracy Comparison

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 2, February 2026

DOI: 10.17148/IJARCCE.2026.15222

System Output Screenshots:

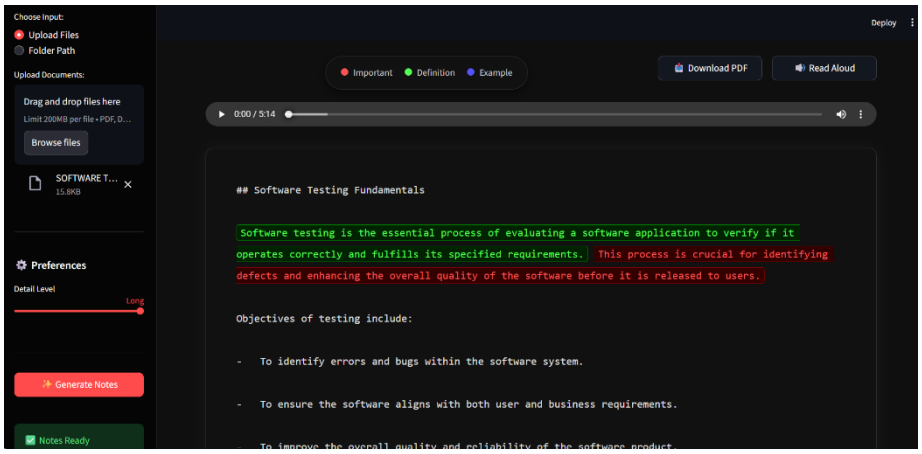


Figure 1.3: Output

VI. CONCLUSION

In conclusion, the Intelligent Multimodal Notes Generation System successfully fulfils its intended purpose of providing an automated, reliable, and efficient note-generation platform. The system achieves high performance, maintains accuracy across different input types, and demonstrates strong potential for future enhancement. The project not only meets its functional and technical objectives but also establishes a scalable framework for developing next-generation AI-powered educational tools. The successful implementation and validation of this system confirm its feasibility, practicality, and long-term relevance in the evolving field of Artificial Intelligence and intelligent documentation systems. The Intelligent Multimodal Notes Generation System provides a strong foundation for further research and development. By integrating advanced AI models, cloud infrastructure, mobile platforms, and collaborative features, the system can evolve into a comprehensive intelligent documentation assistant. The future enhancements discussed demonstrate the long-term scalability, adapt- ability, and commercial viability of the proposed system. Continuous innovation and technological upgrades will ensure that the system remains relevant in the rapidly evolving field of Artificial Intelligence and educational technology.

REFERENCES

[1]. Jurafsky, D., & Martin, J. H. *Speech and Language Processing*. Pearson Education.
[2]. Manning, C. D., Raghavan, P., & Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press.
[3]. Vaswani, A., et al. *Attention Is All You Need*. Advances in Neural Information Processing Systems (NeurIPS), 2017.
[4]. Devlin, J., et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Google AI Research, 2018.
[5]. Goodfellow, I., Bengio, Y., & Courville, A. *Deep Learning*. MIT Press.
[6]. Lin, C.-Y. *ROUGE: A Package for Automatic Evaluation of Summaries*. ACL Workshop on Text Summarization, 2004.
[7]. Microsoft Azure & Google Cloud Documentation – AI and NLP Services.
[8]. OpenAI Research Publications – Natural Language Processing and Transformers.
[9]. IEEE Xplore Digital Library – Articles on AI-based Note Generation Systems.
[10]. ResearchGate – Studies on Multimodal Learning and Automated Content Summarization.

