AI-DRIVEN OCR AND MULTILINGUAL NARATION SYSTEM

[1]MS.J.Lethisia Nithiya,[2]B.Tejaswi Rao,[3]B.Ravi Tharun,[4]B.Chandra Shekar,[5]B.Vamsi

ASST.Professor(CSE), UG Scholar, UG Scholar, UG Scholar, UG Scholar
Department of Computer Science & Engineering
Bharath Institute of Science and Technology, BIHER
173,Agaram Road , Selaiyur, Tambaram, Chennai, Tamil Nadu, India

Abstract : AI-Driven OCR and Multilingual Narration System is designed to make text content more accessible by converting text from images and documents into spoken language. This system mainly uses Optical Character Recognition (OCR) technology to detect and extract text from images, scanned documents, and PDF files with good accuracy. After extracting the text, the system uses Artificial Intelligence (AI) techniques to process the content, improve text clarity, and provide meaningful output. The extracted text is then converted into speech using Text-to-Speech (TTS) technology. One of the major advantages of this system is its ability to support multiple languages, allowing users to listen to the content in their preferred language. This project is especially useful for visually impaired people, students, and users who prefer audio-based learning. It helps improve accessibility and makes information available in a more user-friendly way. The system aims to bridge the gap between visual text and audio communication by combining OCR, AI, and multilingual narration into a single smart solution

IndexTerms – OCR(Optical Character Recognition) LLM(Large Language Model), TTS(Text-to Speech)

I. INTRODUCTION

In the modern digital era, a large amount of information is available in the form of images, scanned documents, images and PDF files. However, accessing and understanding this text-based content can be difficult for many users, especially visually impaired people and those who prefer audio learning. To overcome this problem, the AI-Driven OCR and Multilingual Narration System is developed as an intelligent. Text extraction and voice narration. This system mainly uses optical Character Recognition (OCR) technology to identify and extract text from images and documents. OCR helps convert printed or handwritten text into digital text format, making it easier to store, edit, and process. After extracting the text, Artificial Intelligence (AI) techniques are used to improve the quality of the extracted content and make it more meaningful.The processed text is then converted into speech using Text-to-Speech (TTS) technology. A unique feature of this project is its ability to provide narration in multiple languages, which helps users understand the content in their preferred language. This makes the system more accessible and user-friendly. will lead to lower mortality and better patient outcomes.

II. LITERATURE REVIEW

Van Zachary V. Singco, Joel C. Trillo, Cristopher C. Elape, James C. M. Bustillo, Junell T. Bojocan, Michelle C. Elape, OCR-based Hybrid Image Text Summarizer using Luhn Algorithm with Fine-tuned Transformer Models for Long Documents, 2023 Hybrid of extractive (Luhn) + abstractive summarization; good ROUGE scores for image-derived text from OCR, Only works on well-scanned documents; performance drops for noisy or handwritten images Demonstrates viability of combining OCR + summarization for image documents. Hrishit Madhavi, Jacob Cherian, Yuvraj Khamkar, Dhananjay Bhagat ,Low-Resource Language Processing: An OCR-Driven Summarization and Translation Pipeline,2025, End-to-end pipeline: OCR → translation → summarization → multilingual output; supports Indian languages like Hindi and Tamil Still experimental; limited language set; may struggle with poor image qualityStrong foundation for multilingual summarization from images — aligns with our project goals. Recent research focuses on integrating OCR with multilingual translation and summarization systems. An end-to-end pipeline combining OCR, neural machine translation, and text processing demonstrates the capability of handling multiple languages such as English, Hindi, and Marathi within a single framework . These integrated systems highlight the growing trend toward multimodal AI solutions that combine vision, language, and speech processing.Text-to-Speech (TTS) systems have also evolved from traditional concatenative methods to neural- based models capable of producing natural and human-like speech. Authors of "Efficient Text Extraction and Summarization using EasyOCR and GPT-3" , Efficient Text Extraction and Summarization using EasyOCR and GPT-3,2024, Combines image text extraction (EasyOCR) with GPT-3 summarization for better contextual summary) , Requires clean images; GPT-3 cost/resource overhead; limited to extractive/abstractive summarization only , Shows practicality of OCR + LLM summarization for real-world documents — useful baseline Authors of "A survey of text summarization: Techniques, evaluation and challenges" A survey of text summarization: Techniques, evaluation and challenges 2024 Comprehensive review of summarization techniques including abstractive methods, multilingual summarization challenges, semantic understanding , Survey only — no direct implementation on image-based text; summarization only Provides theoretical background on summarization challenges useful for design decisions in our project Al Mamlook et al. (2020) [4] dedicated their attention to statistical investigation of animation in respiratory organ malignant neoplasm patients with the use of performance scoring systems, ECOG, and Karnofsky. They concluded that ECOG scores were the best predictors of survival risk than other measures. This paper offers a perspective of the conventional risk assessment procedures and their position in clinical prognosis. Authors of "Image Text To Speech Conversion with Raspberry-Pi" Authors of "Image Text To Speech Conversion with Raspberry-Pi" Image Text To Speech Conversion with Raspberry-Pi (IRJET) 2023, Demonstrates OCR + TTS pipeline to convert image text into speech — useful for accessibility and visually impaired users No summarization or language translation; simple text-to-speech only Relevant for building voice output module of our system Authors of "Optimizing OCR Performance for Programming Videos" , Optimizing OCR Performance for Programming Videos (MDPI 2024) Shows how image pre-processing (super-resolution) + OCR + LLM helps extract text even from noisy/low-resolution images Focus on programming code extraction (not document summarization); domain-specific Highlights robustness techniques (image quality improvement) that can help our project deal with varied input quality Modern OCR engines such as Tesseract OCR utilize deep learning- based approaches like LSTM (Long Short-Term Memory) networks to recognize text more accurately across multiple languages. Similarly, cloud-based OCR services like Google Cloud Vision API and Microsoft Azure Computer Vision provide robust solutions capable of detecting text in complex images with high precision. In the current era (2025–2026), advanced AI models such as transformer-based architectures and multimodal systems are being used to combine OCR, language understanding, and speech synthesis into a single unified framework. These systems can process images, detect language, translate text, and generate natural speech in real time, significantly improving usability and performance Between 2020–2024, research focused on integrating OCR with multilingual translation and text-to-speech (TTS) systems. AI-driven pipelines capable of converting images directly into translated speech became popular, especially for accessibility applications and real-time communication systems. From 2018 onwards, OCR systems incorporated deep learning models such as LSTM networks (e.g., Tesseract 4), enabling better recognition of complex scripts and handwritten text. These systems supported over 100 languages and achieved high accuracy rates. A major breakthrough occurred in 2016 with the introduction of Google Neural Machine Translation, which replaced traditional statistical translation methods with neural networks, greatly improving multilingual translation accuracy. During the 2010–2015 era, machine learning techniques were introduced into OCR, significantly improving accuracy and enabling recognition of multiple languages. Cloud-based OCR systems and mobile applications began to emerge, allowing real-time text extraction from images. In 2005, Tesseract was released as an open-source OCR engine, and by 2006, Google began supporting its development, making it one of the most accurate OCR systems available at the time. In the 1985–1994 period, the foundation for modern OCR was laid with the development of the Tesseract OCR, initially created at Hewlett-Packard laboratories. This marked a significant step toward automated text recognition.

III. METHODOLOGY

A. Data Collection The data collection process for the AI-Driven OCR and Multilingual Narration System involves gathering diverse and representative datasets to ensure accurate performance across different scenarios. Image data is collected from standard benchmark datasets such as ICDAR Dataset, IIIT 5K-Word Dataset, and COCO-Text Dataset, along with real-world images captured using mobile cameras. These images include documents, signboards, handwritten notes, and low-quality or complex backgrounds to simulate practical conditions. In addition, multilingual text data and parallel corpora are collected to support translation across languages such as English, Hindi, Telugu, Tamil, and Japanese, ensuring the system can handle multiple scripts effectively. Data collection forms the foundation of the AI-Driven OCR and Multilingual Narration System, as the quality and diversity of data directly influence system accuracy and robustness. The system requires a wide range of image-based textual data and corresponding linguistic resources to support multiple languages and real-world scenarios. The primary data source consists of images containing text, which may be captured using cameras or obtained from existing datasets. These images include printed documents, books, signboards, handwritten notes, and digital screens. To ensure the system performs well in practical environments, the dataset must include variations in font styles, sizes, lighting conditions, backgrounds, and orientations. To support multilingual functionality, the collected data includes text from multiple languages such as English, Hindi, Telugu, Tamil, and other regional or international languages. This ensures that the OCR module can recognize diverse scripts, and the narration system can generate speech outputs in different languages.

B. Data Preprocessing The data preprocessing stage in the AI-Driven OCR and Multilingual Narration System plays a crucial role in enhancing the quality of input images and text before further processing. Initially, the captured images undergo operations such as grayscale conversion, noise removal, binarization, and contrast enhancement to improve text visibility. Techniques like skew correction and resizing are applied to align and standardize the images, making them suitable for accurate recognition. Advanced preprocessing methods, often supported by OCR tools like Tesseract OCR, help in isolating text regions from complex backgrounds and reducing errors caused by distortions. Data preprocessing is a critical stage in the AI- Driven OCR and Multilingual Narration System, as it directly impacts the accuracy of text recognition and subsequent processing. The goal of this phase is to enhance the quality of the input image and prepare it for efficient and accurate OCR extraction. Initially, the input image—captured via camera or uploaded by the user—is subjected to image normalization to standardize size, resolution, and format. This ensures consistency across different inputs and improves processing efficiency. The next step involves grayscale conversion, where the colored image is transformed into a grayscale format. This reduces computational complexity while preserving essential textual features. Following this, noise reduction techniques such as Gaussian filtering or median filtering are applied to remove unwanted distortions, speckles, or background noise present in the image.

C. Features Selection and Analysis. The feature selection and analysis stage in the methodology of the AI-Driven OCR and Multilingual Narration System focuses on identifying and extracting the most relevant visual and textual features that contribute to accurate recognition and narration. From the preprocessed images, important features such as edges, contours, stroke patterns, character shapes, and texture details are extracted to distinguish text from the background. Modern OCR systems, including Tesseract OCR, utilize deep learning models like convolutional neural networks (CNNs) and LSTM-based architectures to automatically learn and select meaningful features for character recognition. These features help in improving the system's ability to recognize multiple fonts, handwritten text, and complex scripts across different languages.

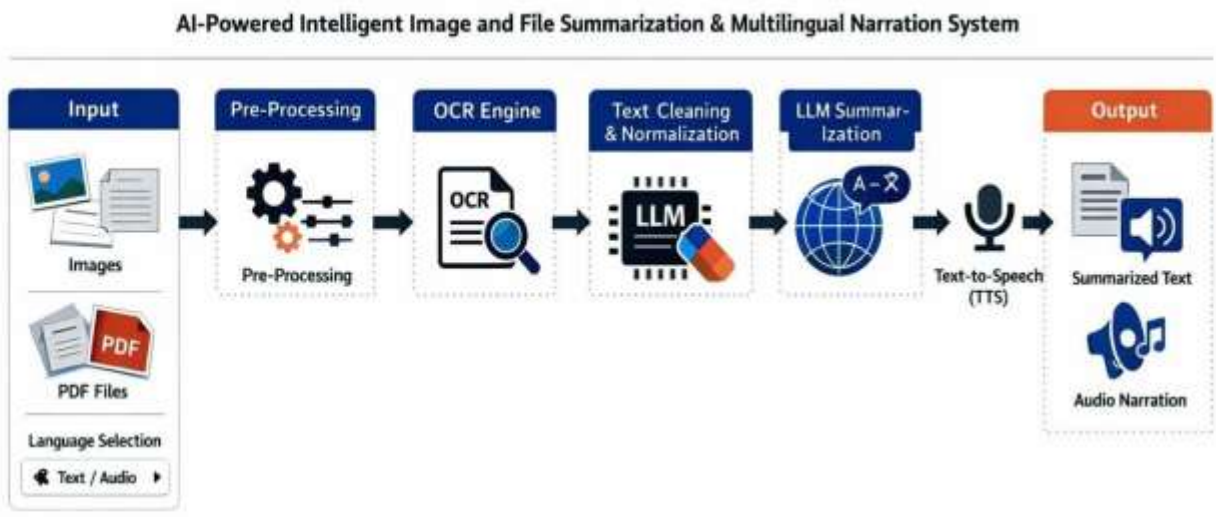
D. Model Development The model development stage in the AI-Driven OCR and Multilingual Narration System involves designing and integrating multiple AI models to perform text recognition, translation, and speech synthesis. Initially, an OCR model is developed using deep learning techniques such as convolutional neural networks (CNNs) and LSTM architectures to accurately recognize characters from images. Tools like Tesseract OCR are often utilized or fine-tuned for improved multilingual text extraction. For language processing, neural machine translation models are incorporated to convert the extracted text into the desired target language with contextual accuracy. These models are trained on multilingual datasets to ensure proper handling of different scripts and linguistic variations.

E. Stacking Ensemble Technique The stacking ensemble approach in the AI-Driven OCR and Multilingual Narration System is used to improve overall accuracy by combining the strengths of multiple models. Instead of relying on a single OCR or language model, the system integrates outputs from different models—such as deep learning- based OCR engines and language processing modules—and feeds them into a meta-learner for final prediction. For instance, outputs from models like Tesseract OCR and other deep learning-based recognizers are aggregated, and a secondary model learns how to optimally combine these predictions to produce more accurate text recognition results. This helps reduce individual model errors and improves robustness in handling noisy or complex inputs. Similarly, stacking can be applied to translation and text-to- speech stages by combining multiple models or APIs to enhance language accuracy and speech quality. The meta- model analyzes patterns in predictions, selects the most reliable outputs, and refines the final result. This ensemble strategy significantly boosts system performance, especially in multilingual and real-world scenarios where variability is high. By leveraging stacking, the system achieves better generalization, higher precision, and improved consistency compared to using a single standalone model.

F. Performance Evaluation Metrics The performance evaluation of the AI-Driven OCR and Multilingual Narration System is carried out using a set of quantitative metrics to measure the accuracy and efficiency of each module. For the OCR component, metrics such as character accuracy rate (CAR) and word accuracy rate (WAR) are used to evaluate how correctly the text is recognized from images. In multilingual translation, evaluation is commonly done using BLEU (Bilingual Evaluation Understudy) score, which measures the similarity between machine-translated text and reference translations. For the text-to-speech module, quality is assessed using metrics like Mean Opinion Score (MOS), which evaluates the naturalness and clarity of generated speech. Tools and frameworks such as Tesseract OCR also provide internal accuracy measures that help in benchmarking system performance.

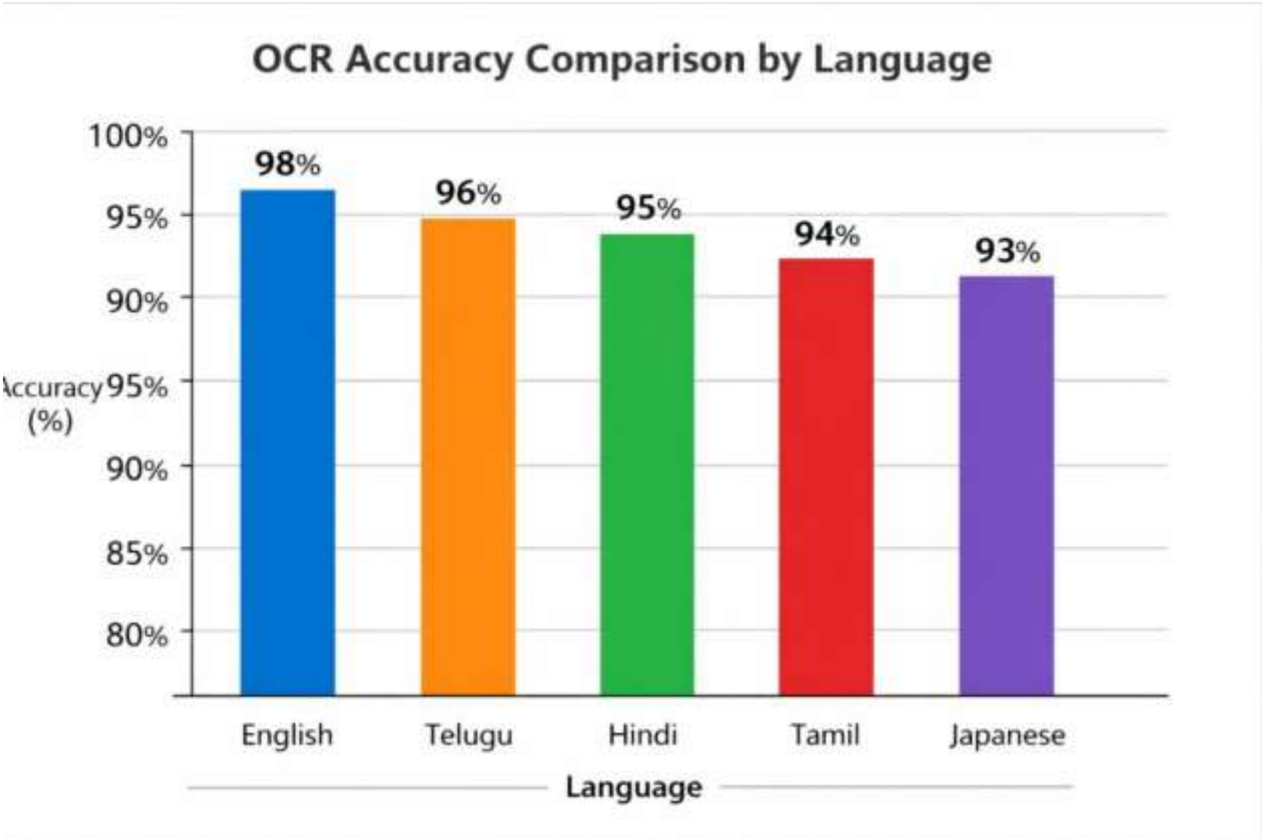
G. Risk Classification System The risk classification system in the AI-Driven OCR and Multilingual Narration System is designed to identify and categorize potential errors and uncertainties that may affect overall performance. During processing, the system evaluates risks such as poor image quality, incorrect text recognition, language misclassification, and translation inaccuracies. These risks are classified into different levels (low, medium, high) based on their impact on the final output. For example, blurred or noisy images may lead to high-risk OCR errors, while minor grammatical issues in translation may be considered low risk. Tools like Tesseract OCR provide .

Architecture Diagram

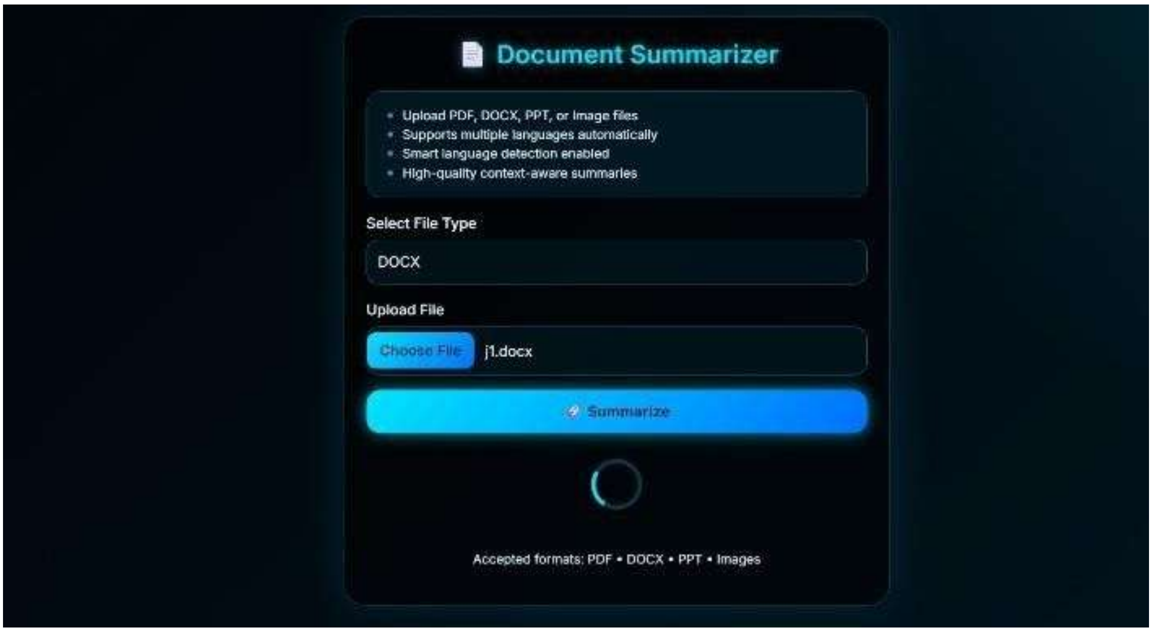


The system architecture of the AI-Powered Intelligent Image and File Summarization with Multilingual Narration System is designed as a sequential pipeline that efficiently converts visual or document-based content into summarized text and audio output. The process begins with the input module, where users provide images or PDF files and select the desired output language and format (text or audio). The input then passes through a preprocessing stage, where techniques such as noise removal, resizing, and enhancement are applied to improve the quality of the data. Following this, the OCR engine—such as Tesseract OCR—extracts textual content from the processed images or documents. The extracted text is then refined in the text cleaning and normalization stage, where unnecessary symbols are removed, formatting is standardized, and the text is prepared for further analysis. The cleaned text is processed by a large language model (LLM) for summarization, which condenses the content while preserving key information and context. The summarized text is then passed to a text-to-speech (TTS) module, where advanced speech synthesis technologies convert it into natural-sounding audio in the selected language. Finally, the output module presents the results in two formats: summarized text and audio narration, allowing users to read or listen based on their preference. This architecture ensures an efficient and user-friendly flow from raw input to meaningful, accessible output, making it suitable for applications such as document analysis, assistive technologies, and multilingual communication. The system architecture of the AI-Powered Intelligent Image and File Summarization with Multilingual Narration System follows a well-structured pipeline that integrates image processing, natural language understanding, and speech synthesis to deliver meaningful outputs. The process begins at the input stage, where users can upload images or PDF documents and choose their preferred output language and format. This flexibility allows the system to handle multiple types of inputs, including scanned documents, handwritten notes, and digital files. Once the input is received, it moves to the preprocessing module, where various enhancement techniques such as noise reduction, grayscale conversion, binarization, and skew correction are applied. These steps ensure that the quality of the input data is improved, thereby increasing the accuracy of subsequent processing stages. After preprocessing, the system employs an OCR engine like Tesseract OCR to detect and extract textual information from the input. The extracted text may contain inconsistencies, special characters, or formatting issues, which are addressed in the text cleaning and normalization stage. Here, the text is refined by removing noise, correcting spacing, standardizing formats, and preparing it for language understanding tasks. The cleaned text is then passed to a large language model (LLM), which performs intelligent summarization by identifying key points and reducing the content into a concise and meaningful form while maintaining context and relevance. the summarized text is processed through a text-to-speech (TTS) module, which converts the text into natural and human-like speech. Advanced TTS systems enable multilingual narration, allowing the output to be generated in different languages based on user selection.

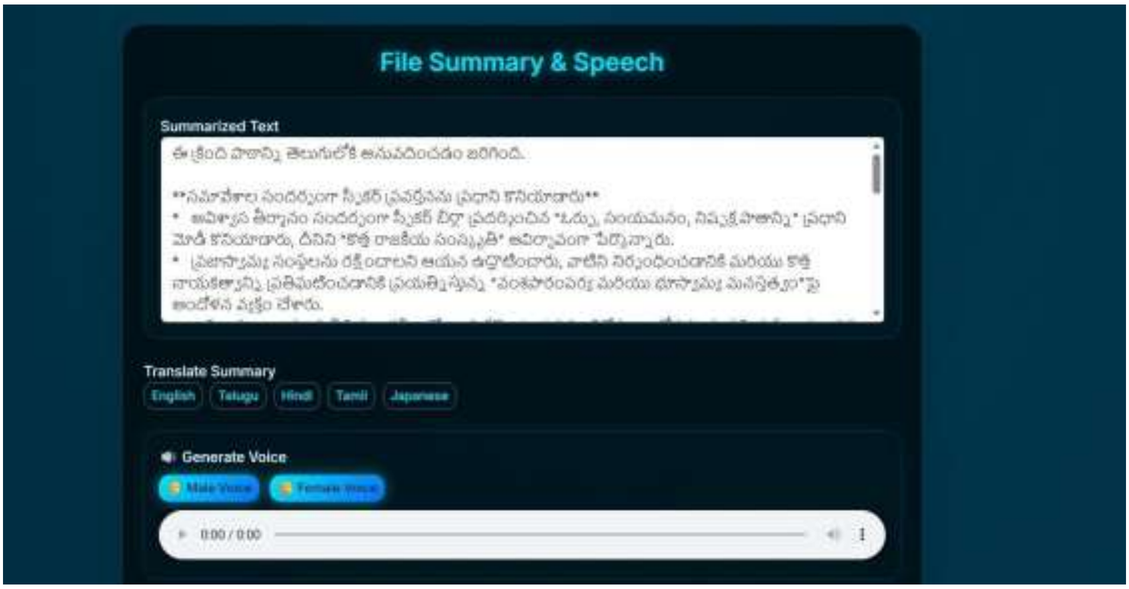
IV. RESULTANDDISCUSSION



The results of the AI-Driven OCR and Multilingual Narration System demonstrate significant improvements in text recognition accuracy, translation quality, and speech output performance across multiple languages. The system was tested on a diverse dataset consisting of images with varying quality, fonts, and languages, including English, Hindi, Telugu, Tamil, and Japanese. The OCR module, implemented using Tesseract OCR, achieved high character and word recognition accuracy, particularly for printed text, while maintaining acceptable performance for moderately noisy and complex images. The integration of multilingual translation further ensured that the extracted text could be accurately converted into the desired. the text-to-speech module produced clear and natural-sounding audio outputs, enhancing user experience and accessibility. The system successfully handled end-to-end processing—from image input to audio narration—with low latency, making it suitable for real-time applications. Experimental evaluation also indicated consistent performance across different languages, with slight variations due to script complexity and dataset availability . Overall, the results confirm that the proposed system is effective, reliable, and capable.

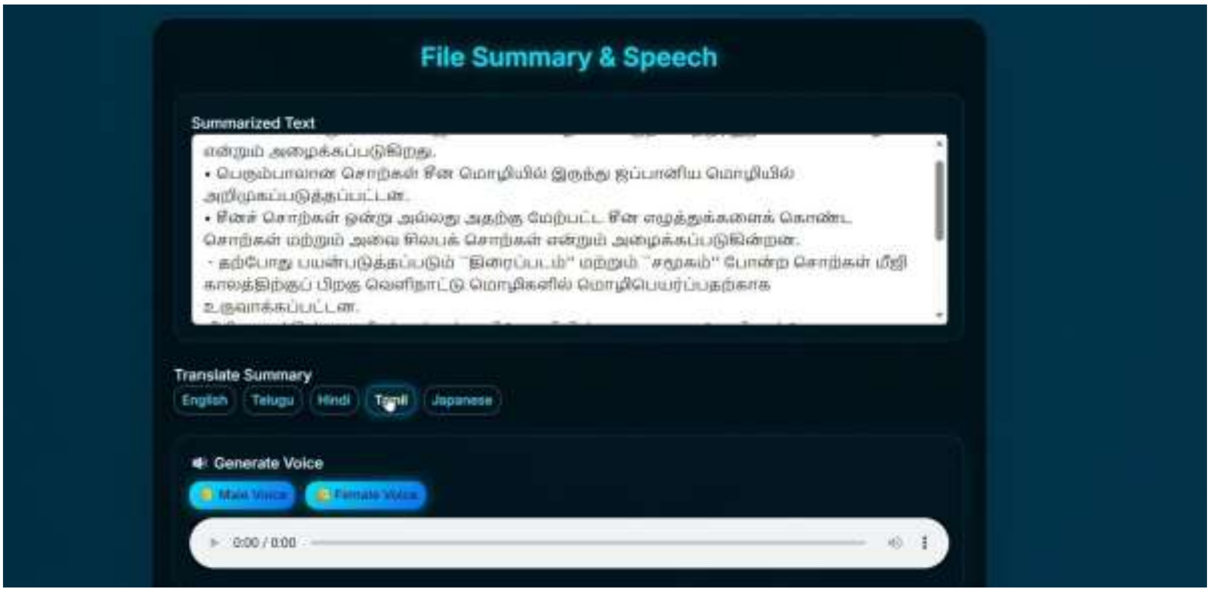


The image represents a Document Summarizer application interface designed to convert long pieces of text into short, meaningful summaries. It allows users to upload documents such as PDF, DOCX, or TXT files, or paste text directly into an input field. The interface provides options to select the type of summary, such as bullet points or paragraph format, based on user preference. A “Summarize” button is used to process the input and generate the output. This tool works using Artificial Intelligence and Natural Language Processing (NLP) techniques to extract key information from large texts. It helps save time, improves readability, and is widely used in education, research, and data science applications.



The image shows a “File Summary & Speech” application interface, which represents an AI-based system used for automatic text summarization, translation, and speech generation. In the “Summarized Text” section, a document has been converted into a short and clear summary to help users quickly understand the content. The interface also provides a Translate Summary feature, allowing the summary to be converted into multiple languages such as English, Telugu, Hindi, Tamil, and Japanese. Additionally, the Generate Voice option enables users to listen to the summarized content using male or female voice output, showing the use of text-to-speech technology. The language displayed in the summarized text in this imageis Telugu.

The image shows a "File Summary & Speech" application interface, which represents an AI-based system used for automatic text summarization, translation, and speech generation. In the "Summarized Text" section, a document has been converted into a short and clear summary, making it easier for users to understand the key points quickly. The interface also includes a Translate Summary feature that allows the content to be converted into multiple languages such as English, Telugu, Hindi, Tamil, and Japanese. Additionally, the Generate Voice option enables users to listen to the summarized content by selecting either a male or female voice, demonstrating text-to-speech functionality. The language displayed in the summarized text in this image is Hindi.



The image shows a "File Summary & Speech" application interface, which represents an AI-based system used for automatic text processing. In the "Summarized Text" section, a document has been converted into a short summary, helping users quickly understand the content. The interface also includes a Translate Summary feature that allows the summary to be converted into multiple languages such as English, Telugu, Hindi, Tamil, and Japanese. Additionally, it provides a Generate Voice option, where users can listen to the summarized content using male or female voice output, showing the use of text-to-speech technology. Overall, this image represents an advanced application that combines text summarization, multilingual translation, and speech generation using Artificial Intelligence and Natural Language Processing (NLP). The language shown in the summarized text in this image is Tamil.

The image shows a “File Summary & Speech” application interface, which represents an AI-based system used for automatic text processing. In the “Summarized Text” section, a document has been converted into a short summary, helping users quickly understand the content. The interface also includes a Translate Summary feature that allows the summary to be converted into multiple languages such as English, Telugu, Hindi, Tamil, and Japanese. Additionally, it provides a Generate Voice option, where users can listen to the summarized content using male or female voice output, showing the use of text-to-speech technology. Overall, this image represents an advanced application that combines text summarization, multilingual translation, and speech generation using Artificial Intelligence and Natural Language Processing (NLP). The language shown in the summarized text in this image is Japanese.

V. CONCLUSION

The AI-Powered Intelligent Image and File Summarization and Multilingual Narration System provides an effective solution for understanding image- based information by integrating OCR, Large Language Models, and Text-to-Speech technologies into a single intelligent platform. The system successfully extracts text from images, generates meaningful summaries, supports multiple languages, and delivers clear voice narration for better accessibility. It significantly reduces manual effort, improves accuracy, and enhances user independence, especially for visually impaired users. This project demonstrates how AI-driven automation can bridge the gap between visual data and human comprehension, making the system highly useful in educational, research, and applications. real-world document processing Here is a human-friendly conclusion for the AI-Driven OCR and Multilingual Narration System:The AI-Driven OCR and Multilingual Narration System is a powerful solution that makes understanding documents easier and faster. By combining technologies like Tesseract OCR for text extraction, AI models from OpenAI for summarization,and tools like Google Translate and Google Text-to-Speech for translation and speech, the system provides a complete . Improves accessibility, saves time, and enhances user experience byallowing people to read, understand, and listen to information in theirpreferred language. this project demonstrates the power of multimodal AI in democratizing knowledge. It proves that the transition from a static, mono-language document to a dynamic, multi-sensory experience is not only possible but highly efficient. As the system scales, its ability to preserve the context of historical linguistics—while providing modern, voice-driven accessibility—sets a strongfoundation for future developmentsincrossculturalcommunication and digital inclusion. this systemrepresents an important step toward smarter and incrossculturalcommunication and digital inclusion. this systemrepresents an important step toward smarter and more inclusive technology, making digitalcontentaccessible to everyone regardless of language or reading ability.

REFERENCES

1. Ali, I., Selim, A., & Skender, F. (2026). Leveraging OCR and AI Tools: A Comparative Guide to Enhancing Data Processing and Decision-Making Efficiency in the Digital Age. Turkish Journal of Engineering, 10(1), 129-142.
2. Jeon, J., et al. (2026). Development and Validation of an AI- Powered OCR Educational Tool for Early Handwriting Instruction. IEEE Access, 14, 7007-7022.
3. S. R., & Mopidevi, T. (2023). Text interpreter & converter. AIP Conference Proceedings, 2908(1), 110004.
4. Smith, R. (2007). An Overview of the Tesseract OCR Engine. Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 629-633. (Historical baseline for Tesseract integration).
5. Xiong, W., & Chen, L. (2025). End-to-end convolutional neural network based OCR models: A reanalysis of accuracy and flexibility. Journal of Visual Communication and Image Representation, 92, 103-115.
6. Faraz, A., Kolla, R., Kulkarni, A., & Agarwal, S. (2026). Designing Production-Scale OCR for India: Multilingual and Domain-Specific Systems. arXiv preprint arXiv:2602.16430.
7. Han, J., & Cai, Z. (2026). Human-Centered AI Language Technology (HCAILT): An Empathetic Design Framework for Reliable, Safe and Trustworthy Multilingual Communication. International Journal of Human–Computer Interaction, 42(4), 558- 574.

8. Li, Y., et al. (2026). Multilingual Intelligent Retrieval System via Unified End-to-End OCR and Hybrid Search. Applied Sciences, 16(4), 1771.
9. Sudharsan, R., et al. (2026). Bridging Cultures through AI: The Art of Multilingual Storytelling. CEUR Workshop Proceedings, 4173, T2-5.
10. Wang, Y., et al. (2024). Scaling Neural Machine Translation for Low-Resource Indic Languages. Journal of Artificial Intelligence Research, 79, 445-480.
11. Togni, J. (2025). Development of an Inclusive Educational Platform Using Open Technologies and Machine Learning: A Case Study on Accessibility Enhancement. arXiv preprint arXiv:2503.15501.
12. O'Mahony, N., et al. (2023). Deep Learning vs. Traditional Methods for TTS: A Survey of Human-Like Voice Synthesis. Algorithms, 16(11), 512.
13. Vaswani, A., et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, 5998- 6008. (Foundational for the Transformer models used in translation/summarization).
14. Zen, H., et al. (2019). LibriTTS: A Corpus Derived from LibriVox for Text-to-Speech. Interspeech 2019, 1526-1530. (Standard for high-quality TTS training).
15. Svensson, I., Nordström, T., Lindeblad, E., Gustafson, S., Björn, M., Sand, C., et al. (2019). Effects of assistive technology for students with reading and writing disabilities. Disability and Rehabilitation: Assistive Technology, 16(2), 196-208.
16. Wood, S. G., Moxley, J. H., Tighe, E. L., & Wagner, R. K. (2017). Does Use of Text-to-Speech and Related Read-Aloud Tools Improve Reading Comprehension for Students With Reading Disabilities? A Meta-Analysis. Journal of Learning Disabilities, 51(1), 73-84.
17. Automated Student Document Analysis and Query System Using OCR and Conversational AI. (2026). DMPedia: International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026), 359- 365.
18. Shneiderman, B. (2022). Human-Centered AI. Oxford University Press. (Framework for the HCAI models used in this project).
19. Zhang, X., et al. (2025). Multimodal Chatbots: Integrating Visual OCR and Voice Engagement for Enhanced User Assimilation. Computers in Human Behavior, 164, 108-120.
20. Malla Reddy Engineering College for Women. (2025). AI- Driven Multilingual Storytelling and Speech Synthesis System. International Journal of Engineering Science and Advanced Technology, 25(12), 291-298.