International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Automating Document Narration: A Deep Learning-Based Speech Captioning System for Visually impaired Person

Pritam Langde[1], Shrinivas Patil[2] Prachi Langde[3]

Submitted: 02/09/2024 **Revised:** 20/10/2024 **Accepted:** 28/10/2024

Abstract

This paper presents an innovative deep learning-based speech captioning system designed to enhance document accessibility for visually impaired individuals. Leveraging a modular architecture that integrates Convolutional Neural Networks (ResNet50), Long Short-Term Memory (LSTM) networks, Optical Character Recognition (OCR), and Text-to-Speech (TTS) technology, the system transforms both textual and visual content from printed documents into real-time, natural-sounding audio. The proposed framework employs image preprocessing and intelligent segmentation techniques to distinguish between text and image regions, followed by content-specific processing—text is extracted via Tesseract OCR, while visual regions are described using an image captioning model based on ResNet-LSTM integration. The summarized content is then converted into speech using the Google TTS API. A custom-built hardware assembly with a mobile-mounted camera, adjustable alignment, and portable design ensures ease of use in real-world settings.

Experimental evaluation on a 100-document dataset demonstrates high accuracy rates— 94% for text recognition, 91% for image detection, and 89% for caption generation. Quality assessments reveal minimal error margins, affirming the system's reliability and effectiveness. This study underscores the potential of AI-driven multimodal solutions in promoting inclusive information access and enabling independent navigation of printed materials by visually impaired users. Future work will focus on real-time enhancements and participatory design inputs from end users to further optimize the system's usability and impact.

Key words : *Assistive Technology, Deep Learning LSTM, Image Captioning Optical Character Recognition (OCR) Text-to-Speech (TTS)*

I. Introduction

In the rapidly advancing technological landscape of the 21st century, education has become an indispensable facet of personal and societal development. However, a significant portion of the global population continues to face formidable barriers to accessing education due to disabilities, particularly visual impairments. According to the World Health Organization (WHO), Vision impairment is a significant global health challenge, affecting at least 2.2 billion people worldwide[1]. Among them, nearly 1 billion cases could have been prevented or remain unaddressed. The primary causes of vision impairment and blindness include refractive errors and cataracts, yet access to effective interventions remains inadequate. Globally, only 36% of individuals with distance vision impairment due to refractive errors and merely 17% of those affected by cataracts receive appropriate treatment. This lack of access not only impacts individual well-being but also contributes to substantial economic losses, with the estimated annual global productivity cost reaching US$ 411 billion[1]. While vision impairment can occur at any age, it predominantly affects individuals over 50 years old.

Despite the efforts of both government and non-

1Research Scholar, Department of Electronics and Telecommunication Engineering, Shivaji University Kolhapur, India , Asst prof Sanjay Bhokare Group of Institutes Miraj , India

2Professor, Head, Department of Electronics and Telecommunication Engineering, DKTE's Textile and Engineering Institute, Research Centre, Ichalkaranji, India

3 Asst Prof , Department of Computer Engineering, SITCOE Engineering College, Yadrav , Ichalkaranji, India.

1 Corresponding Author:

governmental organizations to reduce the prevalence of blindness and improve accessibility through healthcare and awareness programs, the challenges faced by individuals with visual impairments remain substantial. These statistics highlight not only the scale of the issue but also the profound impact of visual impairment on individuals' ability to interact with their environment and participate in daily activities. The lack of access to basic tools for education, employment, and social interaction often exacerbates their marginalization. For blind individuals, the inability to see necessitates reliance on alternative senses and methods for navigation and communication. Unfortunately, many visually impaired individuals face compounded challenges, as visual impairment is frequently associated with other disabilities, further complicating their inclusion in society.

One of the most significant barriers visually impaired individuals face is access to information. In today's world, knowledge is predominantly shared through written text, which remains largely inaccessible to those who cannot see. Historically, oral traditions were the primary means of information dissemination. With the advent of print technology, sighted individuals gained easier access to information, while blind and visually impaired individuals struggled to keep pace. Braille, a tactile writing system developed to provide visually impaired individuals with the ability to read and write, has served as a valuable tool. However, it faces numerous limitations, including the high cost and limited availability of Braille materials, as well as the fact that not all visually impaired individuals are proficient in reading Braille.

Addressing these challenges requires innovative assistive technologies to improve accessibility, enhance quality of life, and support independent living for visually impaired individuals. In response to these challenges, there is an increasing need for innovative solutions that extend beyond traditional Braille. Speech captioning, which converts written or visual information into spoken descriptions, presents a promising alternative. This technology enables blind and visually impaired individuals to access a broader range of materials, from books to multimedia content, by providing descriptive audio that conveys the information contained in visual formats. By embracing such alternatives, we can foster greater inclusion and ensure that visually impaired individuals are better equipped to participate fully in the world around them. This paper explores the potential of speech captioning as a viable tool for improving accessibility and knowledge sharing for visually impaired individuals.

II. Related Work

Advancements in assistive technologies have greatly contributed to improving the quality of life for visually impaired individuals. Among these, object detection, image captioning, text- to-speech conversion, and deep learning techniques have emerged as promising solutions to enhance independence and accessibility. This literature review explores the various technologies and approaches employed to aid visually impaired individuals, with a focus on recent studies.

The reviewed literature collectively demonstrates significant advancements in assistive technologies for the visually impaired, focusing on object detection, navigation, and user- friendly feedback systems. Early solutions like ultrasonic smart sticks and gloves enhanced obstacle detection through vibration and audio alerts, though often limited in range and convenience. More recent systems incorporate computer vision, AI, and deep learning—such as SSD, CNN, TensorFlow, and LSTM—for real-time object and face recognition, distance estimation, and image captioning. Devices like smart glasses, IoT-integrated sticks, and wearable sensor-based systems provide multimodal feedback (vibration, audio), improving independence and safety. Integration of technologies like GPS, GSM, and time-of-flight sensors has enhanced environmental awareness and navigational decision-making. User- cantered designs, hybrid sensor architectures, and encoder-decoder frameworks have proven effective in delivering contextual information and addressing challenges like overhanging or low-height obstacles. Overall, these innovations mark a shift towards more intelligent, accessible, and context-aware assistive devices that empower visually impaired individuals with greater autonomy and mobility[2][3][4][5][6][7][8]

While these technologies are promising, there are inherent challenges that remain. The limitations of current systems often stem from issues such as the need for high computational power, system complexity, and the requirement for specialized hardware. Additionally, not all visually impaired

individuals are familiar with or proficient in advanced assistive technologies, highlighting the need for more intuitive and user-friendly systems.

Recent advances in deep learning have significantly advanced image captioning, particularly for accessibility and domain-specific applications. Early studies adapted CNN–LSTM frameworks to generate Braille descriptions for blind-deaf users, while subsequent work integrated voice-assistance to support visually impaired individuals. More recent methods incorporate attention mechanisms—such as region-guided attention and attribute prediction— to enhance semantic accuracy and capture cultural or technical nuances in specialized images. Additionally, novel decoding strategies like those based on TextGCN have improved the extraction of complex semantic relationships, resulting in more robust and context-aware captions across diverse image domains.[8][9][10] [11]

DEEP neural networks (DNNs) have achieved great success in many real problems such as image understanding and natural language processing. However, DNN models with large

numbers of parameters are hard to be deployed on lightweight edge devices such as smartphones, which restricts the versatility of deep learning (DL). [12]

Recent studies have applied deep learning to solve diverse problems in image captioning and computer vision. One approach uses RNNs to generate image captions and convert them into Braille, aiding blind-deaf users. Another, *CrackVision* application, applies CNNs and transfer learning to accurately detect cracks in concrete, improving infrastructure monitoring. A third study combines CNNs and RNNs with attention mechanisms to enhance the quality of image captions. These works showcase the adaptability of deep learning in accessibility, safety, and vision-language tasks.[13][14] [15]

The reviewed literature underscores the growing integration of deep learning and computer vision in assistive technologies for the visually impaired. From foundational systems utilizing TensorFlow for blind assistance [3] to more sophisticated decision-making algorithms and wearable hardware innovations [4][16][17], research has demonstrated a strong trajectory toward real-time, user-friendly solutions. Recent advancements include camera vision-to-voice object recognition [5] and integrated image description systems like Citizen Cane showcasing practical applications of image captioning in enhancing accessibility. These efforts are further enriched by contributions in image captioning frameworks [5] and domain-specific applications like Thangka and remote sensing [18][11]. Collectively, this body of work reflects a robust and evolving research landscape focused on bridging the gap between artificial intelligence and inclusive design, highlighting both the technological promise and societal value of deep learning in empowering the visually impaired peoples.

The research surveyed in this review highlights notable advancements in assistive technologies designed for individuals with visual impairments. By leveraging object recognition, real-time image captioning, deep learning, and text-to-speech systems, these technologies have significantly improved mobility, access to information, and personal independence. Despite these strides, challenges remain—particularly in terms of cost, accessibility, and ease of use. As such, future research should prioritize refining these systems, broadening their availability, and promoting widespread adoptio.

In conclusion, the literature reflects a rapidly evolving field marked by both significant achievements and ongoing challenges. The continued integration of advanced deep learning methods offers strong potential to address current limitations and create more effective, user- friendly solutions for document captioning. This paper aims to build on the foundation laid by previous research, exploring novel approaches and methodologies to enhance document accessibility for visually impaired individuals.

III. Architectural Design and Implementation Framework

In addressing the challenges faced by visually impaired individuals in accessing and comprehending printed documents, this research proposes a comprehensive deep learning- based approach for speech captioning of documents.

A. System Overview

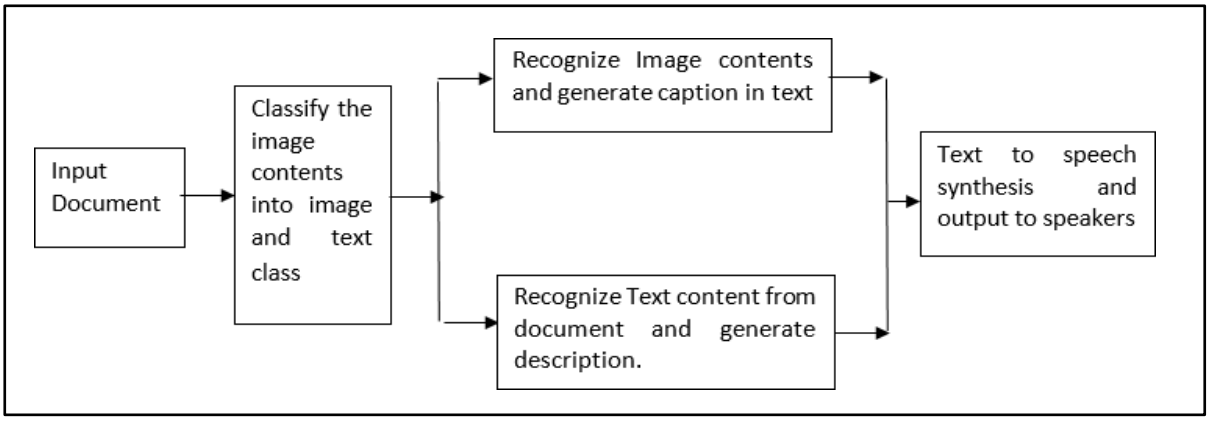

Fig 1 : Conceptual Framework

The system is designed to assist visually impaired individuals by providing detailed audio descriptions of their surroundings or the documents they encounter, as illustrated in Fig. . It integrates image processing, text recognition, and text-to-speech technologies to deliver real- time auditory feedback of visual content. As shown in Fig 1The process begins with capturing an image using a high-resolution camera or document scanner, which may include general text documents, natural image documents, or structured image documents such as historical records. The system first classifies the captured image; if it is a text document, Optical Character Recognition (OCR) is used to extract and digitize the text, which is then summarized into coherent information. For documents containing both text and images, an image segmentation module isolates visual elements for targeted analysis. In the next phase, images within the documents are processed using deep learning models like ResNet50 to recognize objects and generate descriptive captions. These captions, along with the extracted text, are converted into speech using advanced text-to-speech synthesis, ensuring clarity and intelligibility. The synthesized audio is then delivered through high-quality speakers or headphones, enabling users to hear detailed descriptions of their environment or documents in real time. This comprehensive system significantly enhances autonomy by reducing reliance on others and providing timely, accurate auditory information, marking a substantial advancement in accessibility for visually impaired individuals.

B. Research Design Architecture

The system is designed to be a portable and easy-to-use device that visually impaired individuals can use to read printed or handwritten text from images. The architecture of the system consists of several major components, including:

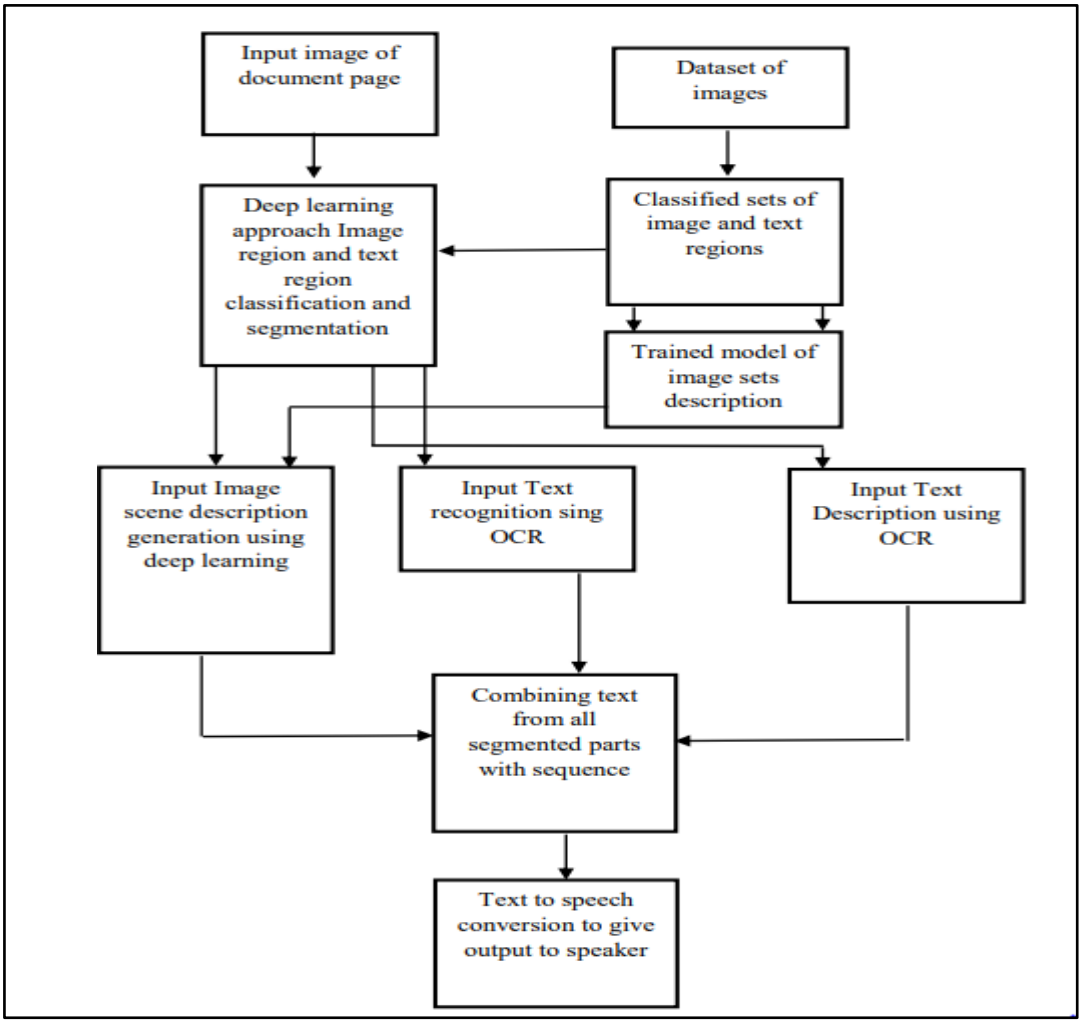

Fig 2: Complete Workflow of the system

As show in in Fig 2. The proposed system is designed to convert both image scenes and textual content from document pages into meaningful speech output. This system integrates deep learning-based image captioning with Optical Character Recognition (OCR) and text-to-speech technologies to create a comprehensive and accessible solution.

The workflow begins with the input of a document page image, which may contain both graphical scenes and textual information. The first processing step involves applying a deep learning-based classification model to identify and segment the image into distinct text and non-text regions. These segmented regions are then classified accordingly.

For image-based regions, a scene description model using deep learning is employed. This module interprets the visual elements and generates natural language descriptions of the scene. For text-based regions, OCR (Optical Character Recognition) is applied to extract readable text. After recognizing text from all segmented parts, the system combines the extracted text in a logical sequence, maintaining reading flow and coherence. The unified textual description is then passed to a text-to-speech (TTS) engine, which converts it into audible speech. The final output is delivered through a speaker interface, providing real-time, intelligible audio feed back to the visually impaired user. This multimodal approach ensures that both visual and textual

contents of the document are effectively communicated, thus enhancing accessibility and promoting independent information consumption.

C. Implementation Framework

The proposed system integrates advanced technologies in image processing, deep learning, and speech synthesis to provide visually impaired individuals with an accessible means of understanding their surroundings and reading printed or digital text. Designed as a real-time assistive solution, the system employs a multi-stage methodology that begins with image acquisition and progresses through preprocessing, text extraction, intelligent content analysis, and audio output generation. At the core of this workflow are components such as Optical Character Recognition (OCR), Convolutional Neural Networks (CNN) including ResNet for robust image classification, and Long Short-Term Memory (LSTM) networks for descriptive image captioning. These modules are supported by a powerful Text-to-Speech (TTS) system, such as the Google TTS API, which converts extracted or interpreted content into natural-sounding speech.

This section outlines the complete design and implementation strategy adopted in the development of the of the system, detailing each phase from image capture to user feedback, and explaining the technical choices made to ensure accuracy, efficiency, and real-time performance.

1. Image Capture

The first step in the process is image acquisition module, where the system utilizes a camera (such as a smartphone camera or a dedicated portable camera) to capture text in real-time. This part of the design requires consideration of the camera's resolution, field of view, and ease of operation. The captured image may contain:

- Pure text documents
- Mixed-content images (text + images)
- Natural scenes or photographs with embedded text

The goal of this stage is to collect visual input in real time, ensuring the quality is sufficient for accurate processing downstream

2. Image Pre-processing

Once the image is captured, preprocessing steps are applied to prepare the image for text extraction. These steps are critical to ensure that OCR operates accurately by removing noise, improving contrast, and isolating text from the background. Key steps include:

- **Grayscale Conversion**: Reduces the image to a single color channel to simplify processing.
- **Noise Reduction**: Removes background clutter using filters like Gaussian blur.
- **Contouring and Image Separation:** Before the actual OCR process, we apply contour detection techniques to identify regions of interest where text is located. Contouring helps isolate the text from complex backgrounds, which can be particularly challenging when reading documents with varying text orientations or noisy backgrounds. We use an edge detection algorithm to find boundaries of the text areas. The system then uses these contours to separate the text from non-text regions, ensuring that only the relevant portions of the image are passed to the OCR engine.
- **Adaptive Thresholding**: It automatically adjusts the threshold value based on the image content, ensuring that text is clearly distinguishable, even in uneven lighting conditions. It Converts images into binary format for better text segmentation.

This stage prepares the image for accurate classification and extraction.

3. Image Classification and Separation

The pre-processed image is passed through a Convolutional Neural Network (CNN), specifically a ResNet (Residual Network) model, for classification and segmentation. This stage determines whether the image primarily contains text, visual content, or a combination of both. It also identifies and segments the regions containing text and those with visual elements. ResNet50, a deep CNN leveraging residual learning, is particularly effective for processing complex, noisy, or low-contrast images—scenarios where traditional OCR techniques may struggle.

4. Image Captioning (ResNet + LSTM)

For the non-text, visual regions of the image (such as figures, photos, or scenes), an Image Captioning

module is applied. A CNN (ResNet) used to extract visual features from the image. ResNet is a deep convolutional neural network known for its ability to handle very deep networks without suffering from the vanishing gradient problem. It works by employing skip connections or residuals to help gradients propagate through the network during training. This architecture is particularly useful for complex image recognition tasks like text extraction from distorted or noisy images

A Long Short-Term Memory (LSTM) network to interpret these features and generate natural language descriptions. To translate the extracted visual features into coherent textual descriptions, a Long Short-Term Memory (LSTM) network is employed. LSTMs, a type of recurrent neural network (RNN), are well-suited for handling sequential data due to their ability to capture long-range dependencies and contextual information over time. In this framework, the high-level visual features—extracted from the preceding CNN-ResNet architecture—are first encoded into a fixed-length representation. This representation serves as the initial input to the LSTM, which then generates descriptive sentences in a step-wise manner. By leveraging the temporal modelling capabilities of LSTM, the system is able to produce fluent and contextually appropriate natural language descriptions that reflect the content of the image. This is especially beneficial for conveying semantic details in images with complex layouts or mixed content, thereby enhancing accessibility for visually impaired users. This pairing (ResNet + LSTM) allows the system to describe images in sentence form (e.g., "A man is sitting at a desk reading a book")

5. OCR Implementation

Optical Character Recognition (OCR) is a crucial component of the system, responsible for converting the pre-processed image into machine-readable text. The OCR process involves the extraction of characters and words from the image, a task that can be complex when the text is in low-quality images or varied fonts. For OCR, we use Tesseract, an open-source OCR engine known for its accuracy and flexibility. Tesseract is capable of recognizing printed text in different fonts, orientations, and sizes. It works by comparing image patterns to its trained data sets of character shapes and then generating the corresponding text. This step produces structured, editable text that can be processed or summarized before speech synthesis.

6. Content Fusion and Text Summarization

The outputs from the OCR engine (text content) and the image captioning module (descriptive captions) are fused. If the content is lengthy, a text summarization algorithm may be employed to condense the information into a concise and meaningful summary. This step ensures that only the most relevant content is passed to the next stage also downstream Text-to-Speech (TTS) synthesis module receives only the most contextually meaningful and user-relevant data, enhancing the overall clarity and usability of the auditory feedback for visually impaired users.

7. Text-to-Speech (TTS) Conversion and Audio output

Once the text has been successfully recognized, the system converts the text into speech. For this task, we use Google Text-to-Speech API, a widely used and reliable service known for its high-quality, natural-sounding voice synthesis. The recognized text is passed to the Google TTS API, which converts it into speech in real-time. The user can select various speech parameters such as pitch, rate, and volume. We ensure that the speed of speech is set at a comfortable pace for most users and offer customization options for individual preferences. The output is played through the device's loudspeaker system or headphones , which is selected to be loud enough to be clearly heard in various environmental conditions. Additionally, the system uses automatic volume adjustment based on ambient noise levels, ensuring that the speech output remains audible to the user.

8. User Interface (UI)

The user interface is designed to be intuitive and accessible for visually impaired individuals. The system's interaction is minimal to ensure ease of use, and auditory feedback plays a central role. The device features a single button for image capture. Once the button is pressed, the system provides an auditory confirmation of the action, and the user is informed when the text is ready to be read aloud. The device offers clear and consistent voice prompts throughout the process.

9. Technologies and Tools Used:

Sr. No.	Component	Technology
1	Image Capture	High-resolution camera/scanner
2	Image Preprocessing	OpenCV
3	Classification	CNN, ResNet50
4	Text Detection & Recognition	Tesseract OCR
5	Image Captioning	CNN + LSTM Architecture
6	Semantic Interpretation	LSTM (Long Short-Term Memory)
7	Text-to-Speech (TTS)	Google Text-to-Speech API
8	Audio Output	Speaker or headphones
9	Programming Language	Python
10	Frameworks & Libraries	TensorFlow, PyTorch, Keras

IV. Experimental Setup and Hardware Configuration

The experimental setup for the proposed deep learning-based speech captioning system is carefully designed to ensure high usability, cost-effectiveness, and real-world applicability for visually impaired users. The integration of modular hardware with an optimized software interface allows for seamless image acquisition, processing, and speech synthesis in an accessible and portable format.

A. Hardware Assembly

The process of acquiring and analysing document images for the purpose of text recognition and speech captioning is an essential element of our system.to full fill the basic need of user we prepared Hardware Assembly. At the heart of this system lies a thoughtfully engineered hardware assembly that emphasizes portability, stability, and ease of use.

Figure 3. Hardware Assembly

Figure 4 : Adjustable height and position for Camera Position in Upward –Downward direction and left right Direction

Figure 5: Foldable Hardware Assembly Design

The physical structure consists of a robust plywood base, measuring 18 inches by 12 inches, providing a stable foundation for mounting components. The framework is constructed using stainless steel rods, selected for their durability, lightweight nature, and corrosion resistance, ensuring longevity and structural integrity even under frequent handling.

To support diverse use scenarios, the system includes an adjustable camera mounting mechanism, which allows movement in upward–downward and left–right directions. This flexibility ensures that users can position the camera to suit varying document sizes, lighting conditions, and ergonomic preferences. The mobile phone camera is centrally mounted and functions as a high-resolution image acquisition device. The adjustable stand enhances alignment accuracy, enabling the capture of clean and legible document images—critical for downstream OCR and captioning tasks.

Fig. 3, Fig 4 and Fig 5 illustrate the hardware architecture, highlighting its foldable design, camera adjustability, and user-friendly assembly, which collectively contribute to the system's mobility and practicality. The entire unit can be disassembled and folded into a compact configuration, making it easy to transport and store. This design is particularly beneficial for applications in home, educational, workplace, and public library environments.

The core components of the hardware setup include:

- Camera Module: A mobile phone with a high-resolution camera is used for document capture. The phone is mounted securely on an adjustable stand, enabling precise control over camera alignment and focus.
- Base Platform: Crafted from plywood, this base provides a non-slip, stable mounting surface for both the stand and the document to be captured.
- Adjustable Stand: Constructed with stainless steel, this stand allows for manual height and angle adjustments to optimize the camera's position relative to the document.
- Host System (Laptop/Desktop): The host system, either a laptop or desktop, handles core tasks such as image processing, OCR, caption generation, and text-to-speech conversion. It requires a minimum configuration of an Intel i3 6th Gen processor, 4 GB RAM, USB connectivity for camera input, and audio output via speakers or Bluetooth headphones. Internet access via Wi-Fi or LAN is recommended for cloud-based TTS services.

The system features a high-resolution mobile camera on an adjustable stainless steel stand with a stable plywood base. Its foldable, lightweight design ensures easy transport and setup. The adjustable mechanism allows precise image capture, improving OCR and caption accuracy. Compatible with standard computers and supporting both wired and wireless audio, it offers a cost-effective, user-friendly, and portable solution for visually impaired users.

5.2 Software Interface

The software interface is a critical component of the proposed speech captioning system, enabling seamless interaction between the hardware and the user. Designed with accessibility and modularity in mind, the interface orchestrates multiple processing stages—from image acquisition to audio output—through an integrated pipeline.

The processing begins with document text recognition, leveraging optical character recognition (OCR) engines to extract readable text from captured images. Simultaneously, image recognition is performed using contour detection following the initial recognition stages, the system employs a hybrid deep learning model comprising Convolutional Neural Networks (CNNs), Residual Networks (ResNet), and Long Short-Term Memory (LSTM) networks. CNNs and ResNet architectures facilitate robust visual feature extraction, even under challenging conditions such as low contrast or noisy backgrounds. LSTM networks are then used for caption generation, effectively modelling temporal dependencies to produce descriptive, context-aware textual outputs. The final output is handled by the Text-to-Speech (TTS) module, which transforms both OCR-extracted text and generated image captions into natural-sounding speech. This audio output is delivered through headphones or speakers, providing an intuitive and inclusive experience for visually impaired users. The software interface supports interaction through simple keyboard or voice commands and is optimized for real-time processing on standard computing systems.

This integrated software design ensures high performance, adaptability, and user accessibility, making it well-suited for visually impaired individuals.

V. Dataset Description

This study employs two distinct datasets to support the development and evaluation of the proposed speech captioning system for visually impaired individuals. The datasets include both a publicly available benchmark dataset for image captioning and a custom dataset curated from educational resources. Together, they enable comprehensive training and testing of various components of the system, including image recognition, caption generation, and speech synthesis.

A. Image Document Dataset Preparation

To evaluate the effectiveness of the proposed speech captioning system, a custom dataset comprising 100 image-based document samples for global scene was curated. Each document in the dataset contains a combination of textual content and embedded images, simulating real- world educational and informational materials. The documents were captured using a high- resolution mobile camera under varied lighting conditions to ensure diversity and robustness in input quality. This dataset served as a critical resource for system testing, performance evaluation, and model tuning, ensuring that the proposed solution effectively handles complex, multimodal content encountered in practical scenarios by visually impaired users

B. Flickr8k Dataset

The Flickr8k dataset is widely recognized in the field of image captioning and is utilized in this study for training and validating deep learning models. It contains approximately 8,000 images, sourced from the Flickr photo-sharing platform, encompassing a wide range of real-world scenes and objects. Each image in the dataset is annotated with five human-generated captions, offering diverse and contextually rich descriptions that enhance model learning and generalization.

The dataset is organized into two primary folders: Flickr8k_Dataset, which contains the image files, and Flickr8k_text, which includes the annotation files. The key file, Flickr8k.token, maps each image to its corresponding captions. The dataset has a total size of roughly 1 GB, and its structure is optimized for use with deep learning frameworks.

In this work, the Flickr8k dataset serves as the foundational resource for training the image captioning pipeline. Convolutional Neural Networks (CNNs) and Residual Networks (ResNet) are used to extract image features, while Long Short-Term Memory (LSTM) networks generate descriptive captions. The high-quality annotations and moderate dataset size make Flickr8k particularly suitable for experimentation and benchmarking of caption generation models.

C. Model Training and Evaluation

For model training, we applied a series of data augmentation techniques, such as image rotation, scaling, and cropping, to simulate challenging visual environments and increase the robustness of the model. The fine-tuned ResNet was then trained on the augmented dataset and integrated into our OCR system. During testing, the enhanced model demonstrated significantly improved performance in recognizing complex and low-quality textual content, leading to more accurate and reliable speech captioning. This improvement is crucial for delivering consistent and meaningful audio feedback in assistive applications designed for visually impaired individuals

VI. Results and Discussion

The experimental results demonstrate that the analysis of sample documents involves three primary components: text recognition, image recognition, and the integration of these outputs with respect to no of words and no of images.

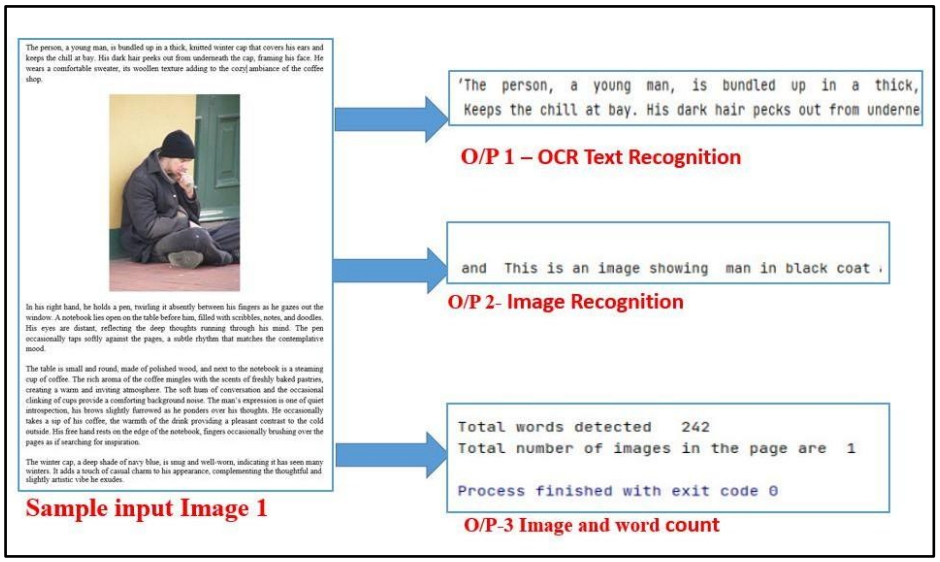

Fig 6. Text , Image recognition and Caption Generation for Sample Document 1

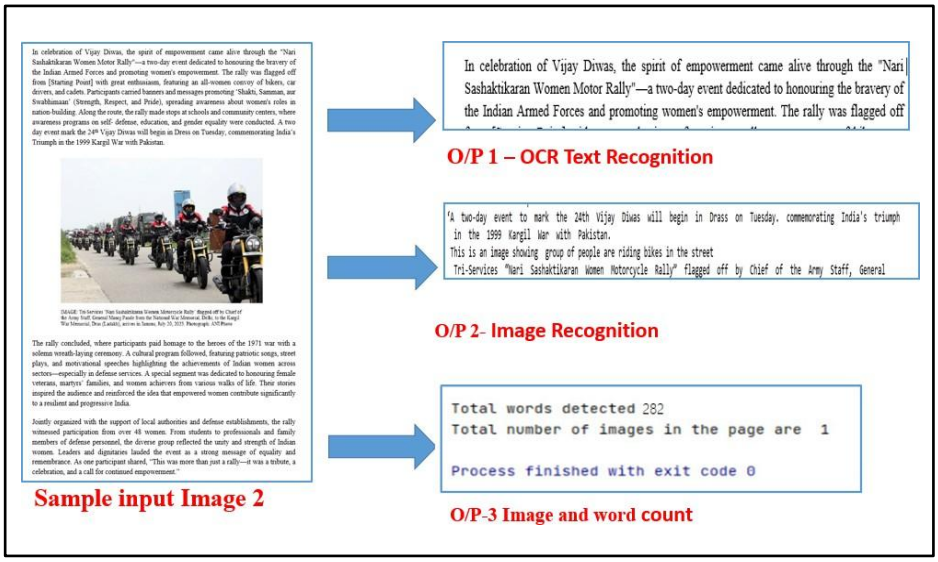

Fig 7 : Text , Image recognition and Caption Generation for Sample Document 2

To enhance text recognition performance, the ResNet model was fine-tuned using the Flickr dataset, which provided a diverse range of textual images for improved generalization. The system's output is categorized into three segments—Output 1 extracts and displays the recognized text, Output 2 identifies and classifies the images within the document, and Output 3 summarizes the total number of words and images detected as shown in Figure 6,7

For evaluation, two sample documents Fig 6,7 were analysed. Sample Document 1 contained 242 words and 1 images, while Sample Document 2 included 282 words and 1 image. These results validate the effectiveness of the Flickr dataset in supporting robust and accurate document analysis, essential to generate meaningful speech captions for visually impaired user

The obtained results are stored in a designated folder as distinct output files in the English language. Both image and text recognition outputs are saved as text files with the ".txt" extension. Specifically, the results from the text recognition process are stored in a file named "MyFile.txt" within the system. These results highlight the effectiveness of the implemented text and image recognition techniques, demonstrating their capability to extract meaningful content from complex document layouts. This significantly aids in the comprehension and analysis of the document content.

Fig 8: Speech Output

The final stage of the process involves generating an audio output fig 8, which can be played through headphones or speakers. This synthesized audio is produced by converting the recognized text and image captions into speech, offering a seamless auditory experience. The purpose of this audio feedback is to convey document information in a clear and accessible format for users, particularly those who are visually impaired. By delivering the output audibly, the system ensures inclusive access to textual and visual information, making it easier for users with diverse needs to interact with printed content in an intuitive and user-friendly manner.

A. Performance Evaluation and result analysis

The performance of our model can be evaluated with separate detection of word and images.

1) Word Detection Performance Evaluation

To evaluate the effectiveness of the proposed system in detecting textual elements from document images, a comprehensive performance analysis was conducted. The evaluation includes both graphical analysis and quality assessment of the detected words. Figure 9 presents a line chart that compares the manually annotated word count with the automatically detected word count for each selected document page within the dataset. The comparison illustrates the system's capability to approximate the actual number of words present on a given page.

Across multiple pages, the detected word count closely aligns with the manual word count, indicating a high level of accuracy in text detection. Although minor deviations are observed in some cases, the overall trend confirms the robustness of the detection module in various layout conditions. In addition to graphical analysis, a quality check test was performed on the detected words to assess recognition fidelity. The results support the conclusion that the system can reliably detect and count words from scanned documents, an essential step toward generating accurate and context-aware speech captions for visually impaired users.

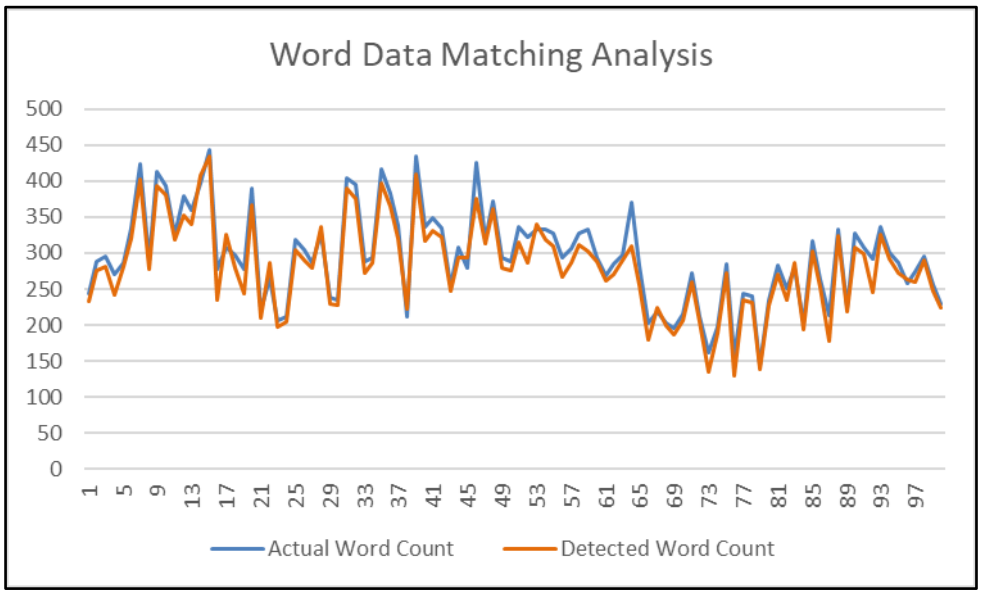

Fig 9 : *Performance evaluation Graph for Text (word) Detection*

- **Quality Check Test for Word Detection**

Once the word detection process for each document is completed, the system's performance is evaluated through a Quality Check Test based on the percentage error between the detected word count and the manually annotated ground truth. This error rate serves as a direct measure of the accuracy and precision of the detection process. A lower percentage error indicates a closer match between the detected and actual word counts, reflecting higher detection fidelity. The evaluation was conducted on a dataset comprising 100 document pages. Results show that 86% of the documents achieved a word detection error rate of less than 7%, demonstrating a high level of accuracy across most of the dataset. 8% of the documents exhibited an error rate between 8% and 15%, suggesting a reasonable degree of accuracy with some potential for refinement. The remaining 6% of documents had an error rate exceeding 15%, indicating more significant discrepancies in word count detection.

These findings confirm that the proposed system is capable of accurately detecting words in a variety of document layouts and quality conditions. The high percentage of documents with low error rates underscores the model's reliability and suitability for downstream tasks such as speech captioning and audio narration for visually impaired users.

2) Image Detection Performance Evaluation

The performance evaluation of the image detection component is carried out through both graphical analysis and a quality check test. Figure 10 presents a line chart that illustrates the comparison between the manually counted and automatically detected number of images for each selected document page within the test dataset. This visual representation highlights the system's ability to identify and count images accurately across a variety of document layouts. In most cases, the detected image count closely approximates the manually verified values, indicating a high level of precision in image detection. These results suggest that the system is effective in locating embedded visual elements within documents, a critical step in enabling meaningful content narration and interaction for visually impaired users.

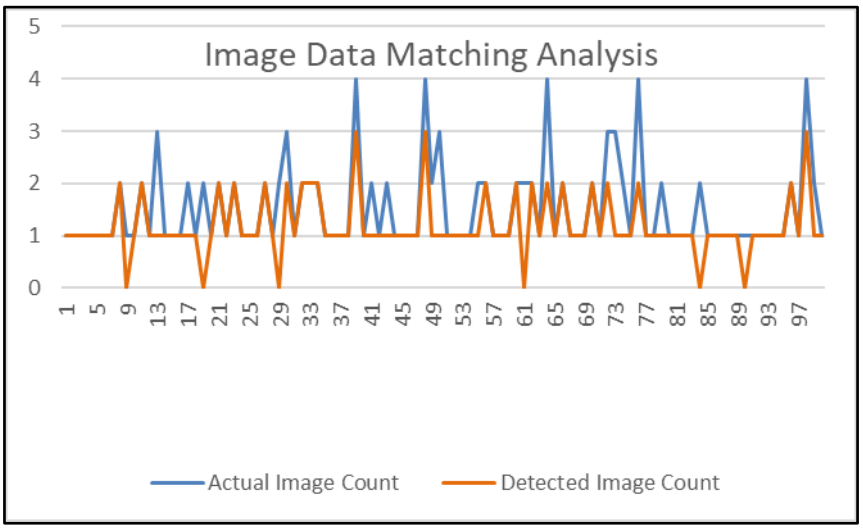

Fig10. Performance Evaluation Graph for Image Detection

- Quality Check Test for Detection of Images:

A critical aspect of evaluating the system's performance involves verifying its ability to accurately detect and count visual elements embedded within document pages. To assess this, a quality check was conducted by comparing the system's detected image count against manually annotated ground truth values across a dataset of 100 documents.

The results demonstrate a strong alignment between the detected and actual number of images in many cases. Specifically, 77% of the documents exhibited no discrepancy between the system's output and manual counts. A smaller subset of documents showed minor deviations, with 14% differing by one image, and 9% showing greater discrepancies.

These findings confirm the system's high reliability in identifying image content, which is crucial for enabling comprehensive and context-aware document narration for visually impaired users. The overall accuracy in image detection supports the integration of this module into broader assistive reading solutions.

3) Caption Detection / Generation Performance Evaluation

The evaluation of caption generation is performed through both graphical analysis and a comprehensive quality check for generated captions. Figure 11 presents a chart comparing the detected image counts with the generated captions for each selected document page.

The results reveal that the caption generation process achieves a high degree of accuracy, with the captions closely aligning with the corresponding images in most instances. This indicates the effectiveness of the captioning mechanism in accurately describing the content of the images, demonstrating its potential to provide meaningful and context-aware captions for visually impaired users.

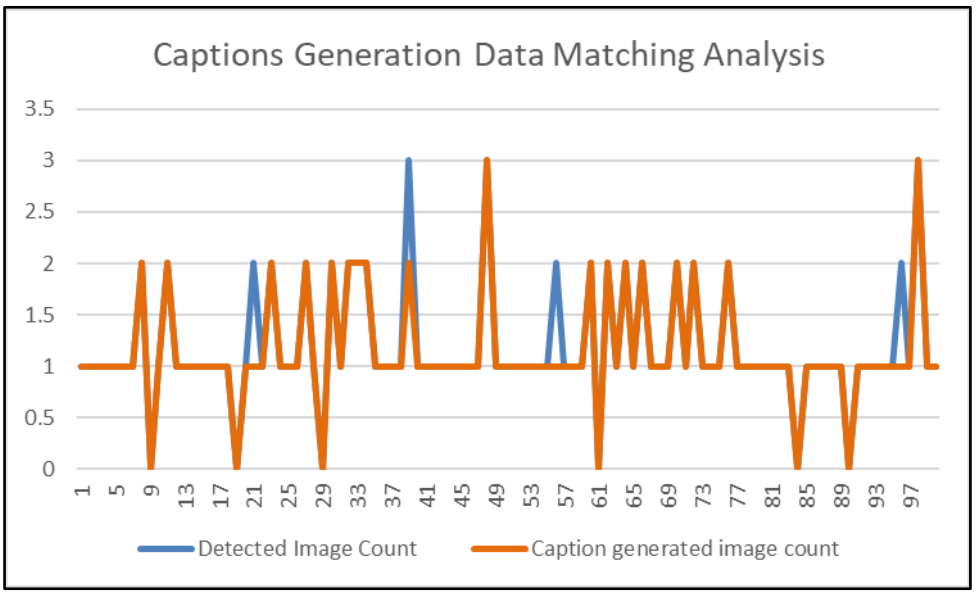

Fig11. Performance evaluation Graph for Caption Generation

- Quality Check Test for Caption Generation

Evaluating the system's ability to generate accurate and contextually appropriate captions for recognized images is a key aspect of assessing overall caption generation performance. To this end, a Quality Check Test was conducted to measure the alignment between automatically generated captions and manually verified references. This evaluation focused on the semantic and contextual relevance of the captions in relation to the detected images.

The results were categorized based on the degree of alignment and accuracy. Captions were classified as: (1) accurate, where the generated caption fully matched the expected description; (2) minor errors, where captions contained slight inaccuracies or omissions but retained general relevance; and (3) irrelevant, where the generated content did not correspond meaningfully to the image.

The analysis was performed on a collection of image-containing documents. The results indicate that 79% of the documents produced accurate captions, 10% included captions with minor errors, and 11% resulted in irrelevant captions. These findings affirm the system's capability to generate reliable and context-aware image descriptions, which is essential for supporting document accessibility and enhancing the user experience for visually impaired readers.

B. Discussion

The evaluation of the proposed system on a curated dataset has delivered promising results, underscoring its effectiveness in accurately recognizing both textual and visual content within documents. Tested on a dataset comprising 100 pages featuring diverse layouts and rich content of both text and images, the system achieved an impressive 94% accuracy in text recognition , 91% in image detection and 89% for Caption Generation , with a tolerance of up to 15% error for text , one-image discrepancy for images and with minor error Captions. These outcomes highlight the system's ability to efficiently process and extract essential information from complex educational material. Overall, the system demonstrated reliable performance in identifying and interpreting printed text and embedded images, paving the way for precise and efficient narration in producing meaningful, content-aware descriptions , an essential step in enhancing the accessibility of educational documents for visually impaired users.

VII. Conclusion and Future Work

This study presents a deep learning-based system for enhancing document accessibility for visually impaired users through automated image-to-text conversion, image detection, and caption generation. Leveraging Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) architectures, the system demonstrated strong performance across key tasks, achieving 94% accuracy in text recognition, 91% in image detection, and 89% accuracy in generating contextually relevant captions. The solution was implemented and evaluated on a global document, highlighting its practical applicability in real-world scenarios.

The integration of these components into a user-centric hardware interface further emphasizes the system's potential as an assistive tool. Future work will focus on refining deep learning models for improved semantic understanding, optimizing the physical design for usability, and incorporating real-time feedback from visually impaired users to drive iterative improvements. These directions aim to strengthen the system's role in promoting digital inclusion and supporting independent access to printed materials.

Conflict of Interest

There is no conflict of interest from author's side.

Data availability statements

The data can be made available on the request to the corresponding authors.

Funding

No Funding has been received for this research work

VIII. BIBLOGRAPHY

[1] WHO, "Blindness and Vision Impairment," *Sustainable Development Goals Series*, 2023.https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual- impairment (accessed Mar. 31, 2024).

[2] H. Akula, G. D. Reddy, and M. Kolla, "Assistive System for the Visually Impaired using Multiple Cameras and Sensors," in *Proceedings of 4th International Conference on Cybernetics, Cognition and Machine Learning Applications, ICCCMLA 2022*, 2022, pp. 363–369. doi: 10.1109/ICCCMLA56841.2022.9989288.

[3] S. C. Jakka, Y. V. Sai, A. Jesudoss, and A. Viji Amutha Mary, "Blind Assistance System using Tensor Flow," *3rd Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2022 - Proc.*, no. Icesc, pp. 1505–1511, 2022, doi: 10.1109/ICESC54411.2022.9885356.

[4] S. Zulaikha Beevi, P. Harish Kumar, S. Harish, and S. J. Lakshan, "Decision Making Algorithm for Blind Navigation Assistance using Deep Learning," in *2022 1st International Conference on Computational Science and Technology, ICCST 2022 - Proceedings*, 2022, pp. 268–272. doi: 10.1109/ICCST55948.2022.10040269.

[5] A. Saicharan, C. Jayalakshmi, B. Sowjanya, K. Raveendra, and M. P. Aslam, "Breaking Boundaries: Advancing Accessibility with Camera Vision to Voice Object Recognition for the Visually Impaired," in *Proceedings of 2nd International Conference on Advancements in Smart, Secure and Intelligent Computing, ASSIC 2024*, 2024. doi: 10.1109/ASSIC60049.2024.10507906.

[6] A. Navhule, Akif, K. Byndoor, A. N. Mohammed Nayaz, D. Shetty, and H. Sarojadevi, "Citizen Cane - An Object Detection and Image Description System for the Visually Impaired," in *Proceedings of the 2024 3rd Edition of IEEE Delhi Section Flagship Conference, DELCON 2024*, 2024. doi: 10.1109/DELCON64804.2024.10866517.

[7] A. Kariri and K. Elleithy, "Astute Support System for Visually Impaired and Blind with Highest Intersection over Union for Object Detection and Recognition with Voice Feedback," in *2024 IEEE Long Island Systems, Applications and Technology Conference, LISAT 2024*, 2024. doi: 10.1109/LISAT63094.2024.10807856.

[8] K. M. Safiya and R. Pandian, "Computer Vision and Voice Assisted Image Captioning Framework for Visually Impaired Individuals using Deep Learning Approach," in *2023 4th IEEE Global Conference for Advancement in Technology, GCAT 2023*, 2023. doi: 10.1109/GCAT59970.2023.10353449.

[9] T. Ghandi, H. Pourreza, and H. Mahyar, "Deep Learning Approaches on Image Captioning: A Review," *ACM Comput. Surv.*, vol. 56, no. 3, Mar. 2023, doi: 10.1145/3617592.

[10] P. Sharma, "Generating Caption From Images Using Flickr Image Dataset," *2024 15th Int. Conf. Comput. Commun. Netw. Technol.*, pp. 1–7, 2024, doi: 10.1109/ICCCNT61001.2024.10724963.

[11] S. Das and R. Sharma, "A TextGCN-Based Decoding Approach for Improving Remote Sensing

Image Captioning," *IEEE Geosci. Remote Sens. Lett.*, 2024, doi: 10.1109/LGRS.2024.3523134.

[12] K. Jivrajani *et al.*, "AIoT-Based Smart Stick for Visually Impaired Person," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, doi: 10.1109/TIM.2022.3227988.

[13] *2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON) : 28th-30th November, FARS Hotel and Resorts, Dhaka, Bangladesh*. IEEE, 2019.

[14] S. Ji, V. Jayaswal, K. Deeksha, S. Kumari, A. Kumar, and P. Bhagat, "A Novel

Approach for Image Captioning using Deep Learning Techniques," in *2024 1st International Conference on Advanced Computing and Emerging Technologies, ACET 2024*, 2024. doi: 10.1109/ACET61898.2024.10730753.

[15] Q. Mohd, I. Hussain, C. V. S. Satyamurty, and R. K. Godi, "Enhancing Accessibility : Image Captioning for Visually Impaired Individuals in the Realm of ECE Advancements," *2024 4th Int. Conf. Technol. Adv. Comput. Sci.*, pp. 317–321, 2024, doi: 10.1109/ICTACS62700.2024.10840791.

[16] S. Samundeswari, V. Lalitha, V. Archana, and K. Sreshta, "OPTICAL CHARACTER RECOGNITION for VISUALL YCHALLENGED PEOPLE with SHOPPING CART

USING AI," *2022 Int. Virtual Conf. Power Eng. Comput. Control Dev. Electr. Veh. Energy Sect. Sustain. Futur. PECCON 2022*, 2022, doi: 10.1109/PECCON55017.2022.9851037.

[17] F. Sen Apu, F. I. Joyti, M. A. U. Anik, M. W. U. Zobayer, A. K. Dey, and S. Sakhawat, "Text and voice to braille translator for blind people," *2021 Int. Conf. Autom. Control Mechatronics Ind. 4.0, ACMI 2021*, vol. 0, no. July, pp. 8–9, 2021, doi: 10.1109/ACMI53878.2021.9528283.