**PUNJAB ENGINEERING COLLEGE (DEEMED TO BE UNIVERSITY)**

Department of Computer Science and Engineering

**MINOR PROJECT**

**REPORT**

**CHAKSHU**

**(Optimized Text-to-Speech and**

**Intelligent Image Captioning Framework)**

Under the Guidance of Dr Manish Kamboj

Punjab Engineering College (Deemed to be University)



Submitted By: -

Raghav Asija (22106010)

Nishtha Mehta (22106037)

Bhavansh Singla (22106043)

Ridhay Neeraj Nathoo (22106053)

Saksham Bhatia (22106056)

**DECLARATION**

We hereby declare that the project work entitled **"CHAKSHU (Optimized Text-to-Speech and Intelligent Image Captioning Framework)"** is an authentic record of our own work carried out at Punjab Engineering College (Deemed to be University), Chandigarh as per the requirements of “Minor Project” for the award of the degree of B.Tech. Computer Science and Engineering, under the guidance and supervision of Dr Manish Kamboj.

We further declare that the information has been collected from genuine & authentic sources and we have not submitted this project report to this or any other university for the award of diploma or degree of certificate examination.

Certified that the above statement made by the students is correct to the best of my knowledge and belief. **Dr Manish Kamboj** (Assistant Professor, Punjab Engineering College)

Date: 04/12/2024

Place: Punjab Engineering College, Chandigarh

**PUNJAB ENGINEERING COLLEGE**

**(DEEMED TO BE UNIVERSITY)**

Department of Computer Science and Engineering



**CERTIFICATE**

It is Certified that the Project work entitled **CHAKSHU: Optimized Text-to-Speech and Intelligent Image Captioning Framework** submitted by **Raghav Asija, Nishtha Mehta, Bhavansh Singla, Ridhay Nathoo** and **Saksham Bhatia** for the fulfilment of Minor Project offered by Punjab Engineering College (Deemed to be University) during the academic year **2024-25** is an original work carried out by the students under my supervision and this work has not framed any basis for the award of and Degree, Diploma or such other titles. All the work related to the project is done by these candidates themselves. The approach towards the subject has been sincere and scientific.

Date: 04.12.2024

Dr Manish Kamboj

CSE Department,

Punjab Engineering College (Deemed to be University)

**ACKNOWLEDGEMENT**

We would like to take this opportunity to thank our college Punjab Engineering College (Deemed to be University), Chandigarh and Department of Computer Science and Engineering (Data Science) for giving us an opportunity to work on this project.

We are immensely grateful to our project mentor Dr Manish Kamboj (Assistant Professor, Department of Computer Science and Engineering) whose continuous guidance, technical support and moral support at times of difficulty helped us to achieve milestones in the given time. She has been a great source of knowledge.

We extend our gratitude to Dr Prabhsimran Singh Bindra for generously sharing his expertise, which significantly contributed to the realization of this research.

We are also thankful to the Committee for evaluation who gave their valuable feedback and guidelines for the betterment and completion of this project.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is a great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Science and Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement several times.

Raghav Asija (22106010)

Nishtha Mehta (22106037)

Bhavansh Singla (22106043)

Ridhay Neeraj Nathoo (22106053)

Saksham Bhatia (22106056)

**ABSTRACT**

**Table of Contents**

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| Figure 01: | Methodology of Image Captioning | 12 |
| Figure 02: | Methodology of Text to Speech | 12 |
| Figure 03: | Open Hand Monument, Chandigarh & its Description on Wikipedia | 18 |
| Figure 04: | Metadata Extracted by the Model | 20 |
| Figure 05: | Caption Generation Function | 21 |
| Figure 06: |  | 14 |
| Figure 07: |  | 16 |
| Figure 08: |  | 16 |
| Figure 09: |  | 17 |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| Table 01: | In Depth Comparison of different TTS Models | 18 |
| Table 02: | Tools & Technologies Used | 20 |

**LIST OF ABBREVIATIONS**

**CHAPTER 1: INTRODUCTION**

The rise of the digital age has transformed how information is shared and accessed, yet a significant portion of the population remains underserved by these advancements. Chakshu, our innovative project, seeks to address this gap by creating a modified version of Wikipedia tailored specifically for the blind and visually impaired. This initiative goes beyond providing basic access; it is a comprehensive solution that enables independence, empowers users with knowledge, and enhances the usability of pre-existing assistive technologies. By integrating cutting-edge tools like text-to-speech conversion and automated image captioning, Chakshu is a step toward creating an inclusive knowledge-sharing ecosystem for everyone.

**1.1 Motivation**

Blindness affects approximately 39 million people worldwide, with an additional 246 million experiencing significant visual impairments [[1]](https://pmc.ncbi.nlm.nih.gov/articles/PMC7253703/#:~:text=The%20estimated%20number%20of%20visually,are%2050%20years%20and%20older.). Despite these staggering numbers, blind individuals face considerable challenges in accessing knowledge that sighted individuals take for granted. While screen readers and other assistive technologies have made progress, they often fall short in offering user-friendly, comprehensive, and customizable solutions. This disparity motivated us to work on Chakshu, a project designed to bridge the gap in accessibility for the visually impaired. Our vision is to create a version of Wikipedia that offers natural-sounding text-to-speech capabilities and descriptive image captions, thereby addressing key barriers to understanding. By empowering independence and delivering a better experience than existing solutions, Chakshu aims to make knowledge more accessible, usable, and impactful for blind users.

**1.2 Problem Statement**

While assistive technologies have evolved, blind and visually impaired individuals continue to face significant obstacles in accessing comprehensive knowledge from platforms like Wikipedia. Current text-to-speech (TTS) systems frequently fall short in handling diverse content formats, such as textual information, figures, equations, and special symbols, making it difficult for users to fully comprehend Wikipedia articles. Similarly, image accessibility remains a critical challenge, as existing solutions often fail to provide accurate and descriptive captions for the wide variety of images present on Wikipedia pages. This limitation restricts users from understanding visual content that is integral to grasping the context of an article.

Moreover, navigating through hyperlinks and interconnected content on Wikipedia is often cumbersome, with limited auditory navigation tools and insufficient support for seamless transitions between pages. In the case of image captioning, the lack of efficient models that balance speed and accuracy further exacerbates the problem, leaving users with incomplete or delayed access to visual information.

**Problem Statement:** Design and develop a system that enables blind and visually impaired users to seamlessly access Wikipedia content, including text, figures, equations, and images, while providing descriptive image captions and intuitive auditory navigation for a comprehensive and inclusive experience.

**1.3 Methodology**

The methodology of the project is organized into two primary components: Text-to-Speech (TTS) and Image Captioning, outlining the step-by-step workflow of our system for delivering a seamless and accessible Wikipedia experience tailored for blind and visually impaired users.

**Text-to-Speech (TTS)**

1. **Review of Existing TTS Systems** - Popular TTS systems, including Coqui TTS, SpeechSynthesis, and pyttsx3, were reviewed for their ability to handle Wikipedia content. Factors such as naturalness of speech, computational efficiency, and symbol/equation rendering were analysed.
2. **Model Selection and Enhancement** - The most compatible TTS model was selected and optimized for Wikipedia's unique content. Enhancements ensured accurate rendering of multilingual text, symbols, and equations, improving overall performance for complex articles.
3. **Interface Development** - A web-based interface was designed to host the TTS model, enabling features such as:
   1. Customization of language, accent, and speech speed.
   2. Real-time API integration for seamless access.
   3. Low-latency audio generation for a smooth user experience.

**Image Captioning**

1. **Review and Shortlisting of Technologies** - Existing image captioning models were evaluated for their ability to describe Wikipedia images. Two complementary models were shortlisted:
   1. LlavaImageCaptioner: Produces detailed captions but requires more processing time.
   2. MetadataImageCaptioner: Generates simple captions quickly using metadata.
2. **Decision Factor Development** - A Python-based decision factor script was developed to dynamically determine which captioning model to use for each image. The process includes:
   1. **Metadata Scraping**: Extracting information such as title and description for each image.
   2. **Evaluation**: Analysing metadata availability, image resolution, and quality to set thresholds for model selection.
   3. **Model Selection**:
      1. If metadata is sufficient, MetadataImageCaptioner is chosen for faster processing.
      2. If metadata is insufficient or the image requires detailed analysis, LlavaImageCaptioner is used.
3. **Output Generation** - The system generates descriptive captions tailored to each image, either through AI-based captioning or metadata processing, ensuring accessibility for all visual content.

By combining these methodologies, the project delivers an integrated solution for rendering both textual and visual content, making Wikipedia more accessible and empowering for blind users.

**Review of existing TTS Systems**

**Model Selection and Enhancement**

**Interface Development**

**Review and Shortlisting of Technologies**

**Decision Factor Development**

**Output Generation**

**Figure 1: Methodology of Image Captioning Figure 2: Methodology of Text to Speech**

**CHAPTER 2: BACKGROUND**

**2.1 Wikipedia and Its Challenges for Accessibility**

Wikipedia, a vast repository of knowledge, operates on the principle of free access to information. However, its extensive reliance on visual and text-based content creates significant accessibility barriers for the visually impaired. Challenges arise from complex page layouts, reliance on images without adequate alternative text (alt-text), and the inability of screen readers to accurately interpret mathematical equations, symbols, or detailed charts. Furthermore, Wikipedia's hyperlink-heavy navigation can overwhelm users relying on assistive technologies, leading to difficulty in accessing and navigating relevant content efficiently.

Statistics highlight the gravity of this issue: globally, around 253 million individuals live with visual impairments, of which 36 million are completely blind. The digital divide between these individuals and sighted users is exacerbated by such accessibility challenges, restricting access to vital educational and informational resources [​[2]](https://www.visioncenter.org/resources/visual-impairment-accessibility/)[[3]](https://easychair.org/publications/preprint/S56B)

**2.2 Information Access for the Visually Impaired in the Digital Age**

Historically, individuals with visual impairments relied on alternative formats like braille, large print, and audio recordings, which were often limited in availability and costly to produce. For instance, only 5% of books in the UK are available in accessible formats, a statistic highlighting the global disparity in information access [(Harris, 2005).](m%20http/www.bpm.Nb-online.org/chapter5html)

Advancements in ICT and assistive technologies have transformed this landscape. Text-to-speech (TTS) systems and braille displays allow direct interaction with digital content, enabling visually impaired individuals to access original materials without intermediaries [(Golub, 2002).](https://www.researchgate.net/publication/39163158_Digital_libraries_and_the_blind_and_visually_impaired) These innovations have brought about significant changes in libraries, schools, and personal use environments, with libraries using enhanced braille book production and digital collections like Bookshare and Project Gutenberg to broaden access.

However, challenges persist. Existing solutions often fall short in handling diverse content formats. For example:

1) Text-to-Speech (TTS) systems, while efficient, often distort nuanced content or fail to adapt to user preferences such as varying speech speeds or accents. They also face challenges with multilingual support or recognizing special characters.

2) Image Captioning Tools struggle with generating meaningful captions for images embedded in complex contexts. Some rely heavily on metadata, leading to oversimplified descriptions, while others demand computationally expensive models for deeper analysis.[[4]](https://www.visioncenter.org/resources/visual-impairment-accessibility/)

Additionally, poorly designed websites often hinder accessibility, emphasizing the need for adherence to standards like WCAG for universal usability [(Williamson, Schauder, and Bow, 2000](https://informationr.net/ir/5-4/paper79.html)).

By leveraging these advancements, projects like Chakshu aim to further improve information accessibility, combining cutting-edge tools with a focus on user needs to bridge remaining gaps.

**2.3 Introduction to Accessibility Technologies and Existing Solutions**

1. **Screen Readers**: Programs like JAWS, NVDA, and VoiceOver interpret on-screen content, converting it into synthetic speech or braille for user interaction. However, their effectiveness depends heavily on website and software compatibility.
2. **Optical Character Recognition (OCR) Systems**: OCR converts printed text into electronic formats that can be read aloud by assistive devices. While effective for high-quality prints, these systems struggle with handwritten content and degraded text quality.
3. **Enhanced Image Magnification Devices**: Tools like closed-circuit televisions (CCTVs) provide magnification for printed materials. Recent advancements incorporate high-resolution displays and portable designs, offering greater flexibility for users.
4. **Braille Displays and Note-Takers**: Braille technology provides tactile access to digital content. Devices like the Perkins Brailler allow users to read and write through raised-dot formatting. Despite their benefits, high costs remain a significant barrier.
5. **Text-to-Speech (TTS) Synthesizers**: These systems convert written text into audible speech, a critical feature for accessing digital content.
6. While these technologies offer numerous benefits, challenges such as limited availability, high costs, and literacy barriers for users persist. Many visually impaired individuals rely on what resources are available rather than accessing the specific information they desire [(Adetoro, 2009).](Adetoro,%20A.A.%20(2009).Relationship%20among%20reading%20interest,%20information%20materials%20availability)

**CHAPTER 3: PROPOSED WORK**

Our work followed a systematic approach to develop an accessible platform for visually impaired users, focused on enhancing Wikipedia content through speech and image description. We began by evaluating and selecting a Text-to-Speech (TTS) system capable of processing multilingual content, mathematical expressions, and multimodal elements. To ensure image accessibility, advanced AI-driven and metadata-based captioning models were integrated, dynamically switching based on image complexity and metadata availability. Wikipedia content was extracted and processed using web scraping and API integration, ensuring real-time updates and embedding descriptive captions. The final system was designed with user accessibility in mind, adhering to WCAG standards and incorporating customizable TTS settings for an optimized user experience.

**3.1 TTS System Selection**

When selecting a Text-to-Speech (TTS) system for the project, several parameters must be carefully evaluated to ensure optimal performance and accessibility.

**System Compatibility -** The TTS system must seamlessly process Wikipedia content, including text, special characters (e.g., mathematical symbols, superscripts, and subscripts), and table data. Compatibility with existing frameworks in the project, such as dynamic captioning models and the user interface, is crucial to provide an integrated experience.

**Resource Efficiency -** The system must align with the project's computational constraints. Factors like low memory usage, efficient processing of large datasets, and reduced latency are essential for hardware with limited resources. For instance, models optimized for edge devices or cloud environments should be evaluated based on their hardware requirements and scalability.

**Speech Customizability -**Adjustable speech features, including:

* **Pace:** Enabling users to control how quickly the TTS reads the content.
* **Pitch and Tone:** Adapting to user preferences, especially for visually impaired individuals who may rely on distinct audio nuances for better comprehension.

**Multilingual and Multimodal Capabilities -** Wikipedia’s multilingual nature requires a TTS system supporting multiple languages. The ability to switch languages dynamically within a single document (e.g., quotes or foreign terms) is essential. Furthermore, the system should handle multimodal content such as equations and charts in a descriptive manner.

**3.2 Design and Development of a Dynamic User Interface**

The proposed system includes a user-centric front-end that ensures accessibility and ease of use for visually impaired users. Key features include:

**Customization Options -** The interface will allow users to adjust TTS settings dynamically, such as speed, volume, and pitch. This ensures a personalized experience tailored to individual preferences.

**Real-Time Responsiveness -** The interface will feature dynamic updates, allowing for real-time feedback. For example, users can navigate through Wikipedia articles with keyboard shortcuts, and the system responds by reading out the text or describing images without delay.

**Accessibility Standards -** Adhering to WCAG 2.1 (Web Content Accessibility Guidelines), the interface will include features like:

* High-contrast visual elements for partially sighted users.
* Keyboard-only navigation support.
* Screen reader-friendly design with ARIA (Accessible Rich Internet Applications) roles.

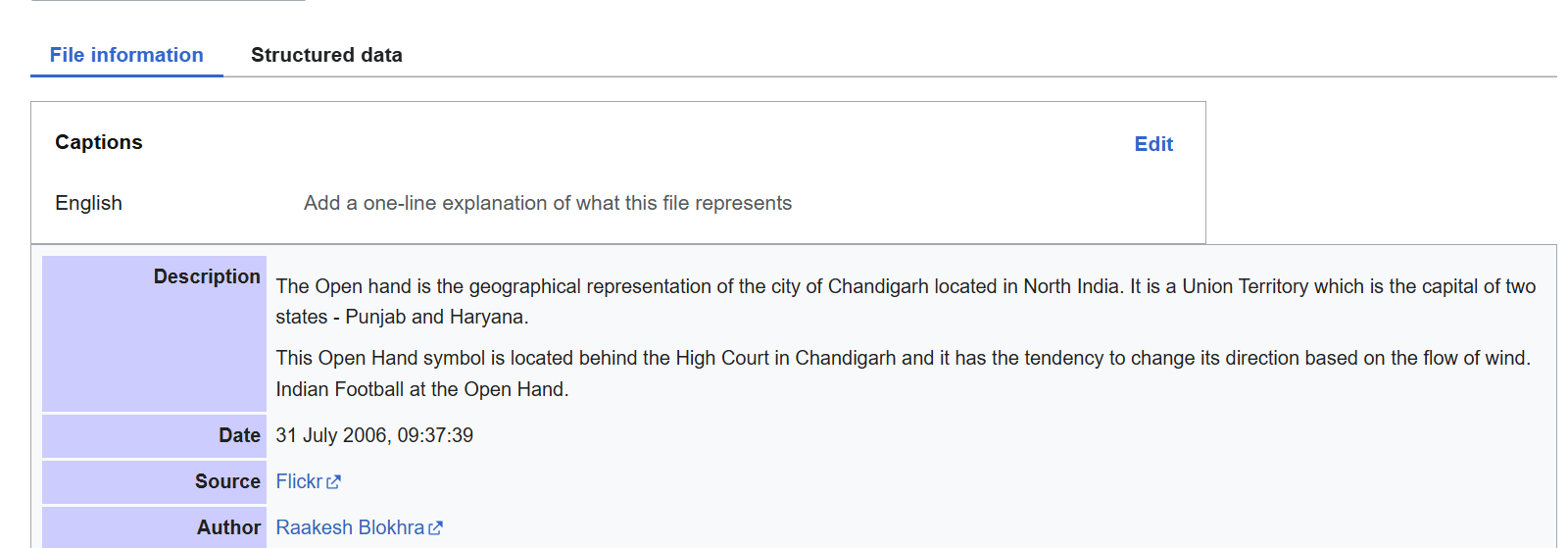
**Prototyping and Testing -** UI prototypes will undergo usability testing with feedback from actual visually impaired users to ensure functionality and comfort.

**3.3 Intelligent Captioning System**

An integral feature of the project is the **dynamic selection of captioning models** for images, ensuring accurate and relevant descriptions.

**Metadata-Based Captioning -** The system first evaluates the metadata associated with an image (title, description, etc.). If metadata is sufficient and descriptive, a lightweight captioning model, such as MetadataImageCaptioner, will be utilized for efficiency.





**Figure 3: Open Hand Monument, Chandigarh & its Description on Wikipedia**

**AI-Driven Captioning -** For images with high complexity (e.g., scientific diagrams, charts, or low metadata availability), advanced AI models like LlavaImageCaptioner are deployed. These models generate rich and descriptive captions based on image resolution, content, and contextual complexity.

**Model Selection Algorithm -** The tool uses thresholds, such as:

* **Metadata Length**: Determines whether metadata provides sufficient detail.
* **Image Resolution**: High-resolution images (>1600x1600 pixels) default to AI-based captioning.
* **Context Complexity**: Analyses keywords in the metadata to identify complex visuals.

**CHAPTER 4: IMPLEMENTATION DETAILS**

**4.1 Text-to-Speech (TTS) System Evaluation and Selection**

To select the most suitable TTS system, we evaluated multiple systems based on factors like efficiency, resource requirements, compatibility, and feature set. The following table summarizes the comparison:

|  |  |  |
| --- | --- | --- |
| **TTS System** | **Advantages** | **Disadvantages** |
| Coqui TTS | Multilingual support, voice cloning, multi speaker features, open-source. | High hardware requirements, slower inference, lacks customer support for enterprise users. |
| Speechbrain (Tacotron2 + HiFi-GAN) | High-quality audio, handles complex text inputs, generates Mel-spectrograms for better fidelity. | Computationally expensive, slower inference, and requires a robust dataset for generalization. |
| Pyttsx3 | Minimal dependencies, offline functionality, lightweight, platform-independent. | Less natural voice output, limited for large-scale deployments. |
| Speech Synthesis API | Lightweight, customizable pitch/pace/volume, cross-browser compatibility, free, no API keys required. | Limited in comparison to advanced neural TTS systems in naturalness and multilingual support. |

**Table 1: In Depth Comparison of different TTS Models**

Selected TTS System: Speech Synthesis API

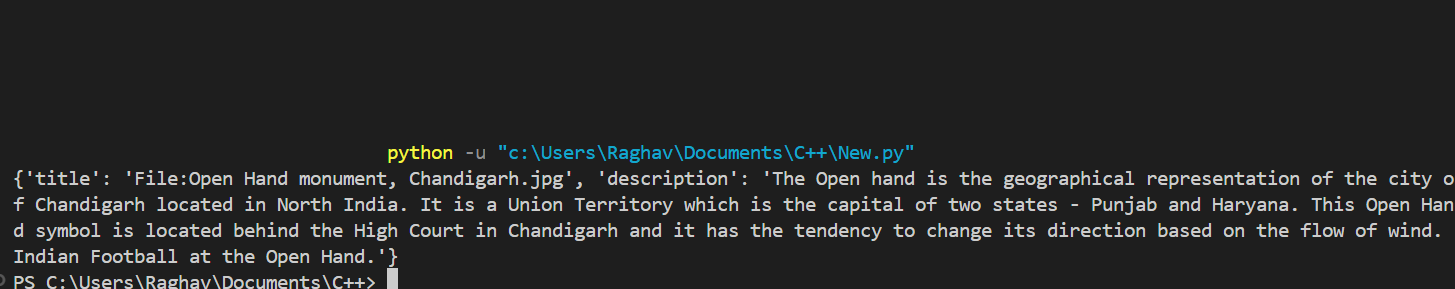
We selected the Speech Synthesis API for its lightweight nature, ease of integration with the frontend, and system-level compatibility with multiple voices and languages. Its support for customizable parameters such as pitch, pace, and volume aligned well with our project's goals of accessibility and user-friendly design.

**4.2 Image Captioning Implementation**

**Dynamic Captioning Tool Workflow**

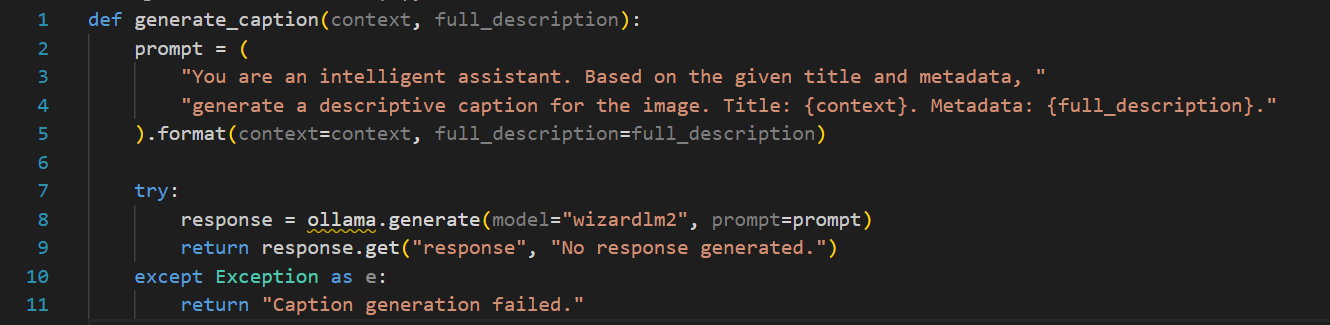
The image captioning system integrates dynamic model selection based on metadata and image complexity. Below is an overview of the primary steps:

1. **Metadata Extraction:** Metadata (title, description) for images is scraped from Wikipedia or Wikimedia. For example:



**Figure 4: Metadata extracted by the model**

1. **Dynamic Model Selection:** The tool uses heuristics such as metadata completeness, image resolution, and context complexity to select between:
   1. Llava Image Captioner: Handles high-resolution or contextually complex images.
   2. Metadata Captioner: Efficient for simpler cases where metadata provides sufficient detail.
2. **Efficiency vs. Time Trade-off** - This trade-off ensures a balance between quality and performance, dynamically adapting to the user's needs.
   1. Llava Image Captioner: Generates richer captions but has higher computational costs and processing time.
   2. Metadata Captioner: Provides faster results but relies on metadata quality.
3. **Caption generation -** This block generates descriptive captions for images using metadata and an AI-based large language model.



**Figure 5: Caption Generation Function**

**4.4 Tools and Technologies**

The development stack for the project included the following:

|  |  |
| --- | --- |
| Category | Technologies Used |
| Programming Languages | Python (backend + model selection script), JavaScript, HTML, CSS (frontend) |
| Frameworks | Flask, React |
| Libraries | BeautifulSoup, Pyttsx3, Ollama, PIL |
| APIs | Speech Synthesis API, Wikipedia API |
| Models | Llava Image Captioner, Metadata Image Captioner |
| Tools | Jupyter Notebook, Postman, Git |

**Table 2: Tools and Technologies Used**

This comprehensive implementation ensures that the system is robust, scalable, and accessible to visually impaired users. The combination of modern technologies and innovative design makes our solution both practical and impactful

**CHAPTER 5: RESULTS AND DISCUSSION**

**5.1 Evaluation Metrics**

**Accuracy:** Accuracy represents the number of correctly classified data instances over the total number of data instances.

**Precision:** Precision is a metric that gives you the proportion of true positives to the amount of total positives that the model predicts.

**Recall:** Number of data points correctly predicted as true out of all the data points that should be predicted as true.

**F1 Score:** F1 Score is a measure that combines recall and precision. There is a trade-off between precision and recall and F1 score can be used to measure how effectively our models make that trade-off.

**5.2 Model Training**

The model was trained on the given dataset and trained for 40 epochs for preprocessed sets C1, C2 and C3 and for 20 epochs for sets C4, C5 and C6 respectively. A snapshot of epoch wise performance progress for set C6 on model M3 is shown below in figure:



The model was then tested on a set of radiographs from the testing dataset and the classification results of the given images are displayed below. The accuracy on the training set was 98%, on validation data it was 77% and 76% on test data.

**Figure 18 :** Results after classification of Radiographs

**5.4 Performance Analysis**

Heatmaps are a method of representing data graphically where values are depicted by color, making it easy to visualize complex data and understand it at a glance. A 5\*5 confusion matrix (5 grades of KL grading scale), is displayed below. Train-Test accuracy over multiple epochs is a technique to get an understanding of how the model is working and performing over the dataset. The line graph for the same is displayed below.

**CHAPTER 6: CONCLUSION AND FUTURE WORK**

**REFERENCES**