iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** III **Month of publication:** March 2025

DOI: https://doi.org/10.22214/ijraset.2025.68030

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

DEEPDETECT: A Text and Object Detection Application

Pranjali S. Marodkar, Om S. Bingule [2], Anuja V. Deulkar [3], Tejas M. Wakade[4], Dr. Aditya P. Bakshi[5]
[1, 2, 3, 4]*Students,* [5] *Professor, Dept of Computer Science and Engineering, Jawaharlal Darda Institute of Engineering and Technology, Yavatmal*

Abstract: ***This project presents an innovative Android application that combines advanced image recognition and multilingual text-to-speech (TTS) conversion technologies to provide an inclusive and efficient user experience. The app enables users to upload or capture images, which are processed using a large language model (LLM) to identify and extract the content from the image. The extracted content is then converted into text in multiple languages (e.g., Hindi, English, Bengali, French, etc.), ensuring accessibility across diverse linguistic groups. Additionally, the text content is passed to the OpenTTS backend to generate speech output in a user-selected language, facilitating auditory content delivery. Users can download the translated text as individual document files (.docx or .doc) in their preferred languages for offline use. The application bridges the gap between visual content, textual interpretation, and multilingual auditory output, making it a valuable tool for education, accessibility, and global communication.***

Keywords: ***Image Recognition, Multilingual Text-to-Speech, OpenTTS, Document Translation, Android Application, Large Language Model (LLM), Accessibility, Multilingual Document Generation***

I. INTRODUCTION

In today's interconnected world, the need for seamless communication across language barriers has become more crucial than ever. This project introduces an innovative Android application to bridge the gap between image recognition, multilingual accessibility, and text-to-speech technology. The app leverages advanced artificial intelligence models to extract and interpret content from images, enabling users to gain insights into their preferred languages. The core functionality begins with the user uploading an image, either captured through the camera or selected from the gallery. A state-of-the-art large language model (LLM) integrated with image recognition capabilities processes the image to identify and extract meaningful content. This extracted text is then passed to OpenTTS, a robust text-to-speech (TTS) framework, which converts the content into speech in multiple languages, including Hindi, English, Tamil, and French. Additionally, the application generates text translations of the content in these languages, storing them as separate files for easy access. The app provides users with the option to download the translated content as document files (e.g., DOC or DOCX), offering greater flexibility for offline use. This comprehensive approach not only ensures accessibility for users across diverse linguistic backgrounds but also caters to those with visual impairments, as the TTS functionality makes the content audible. The backend system, powered by OpenTTS and advanced LLMs, ensures efficient image processing, accurate recognition, and natural-sounding multilingual TTS. The Android app acts as a user-friendly interface, facilitating effortless interaction and providing a holistic experience in accessing, converting, and utilizing content in various formats and languages. This project is a step towards making technology more inclusive, enhancing communication, and enabling users to engage with information in their preferred formats and languages. Text Extraction from Text-Based Image is an Android application that aims to allow the user to extract the text from the image and after extracting if the user wants to translate it into another language he will also translate another language and hear a text contained in a picture that has been taken with a mobile phone. It is an application meant to help those who cannot read a text they encounter, like non-native speakers, the visually impaired, and the blind people, which was estimated at 285 million in 2010 by the World Health Organization.[5].

A. Analysis Of Problem

Language barriers and accessibility challenges often limit the ability of individuals to interact with textual content found in images. This issue is particularly significant for visually impaired individuals, non-native speakers, and those unfamiliar with a particular language. Existing solutions may not provide a seamless integration of image recognition, multilingual text conversion, and text-to-speech (TTS) functionalities. Users require an efficient, inclusive, and easy-to-use application that can process image-based text, translate it into multiple languages, and generate speech output for enhanced accessibility[3].

B. Objective

1) Develop an Android application that integrates image recognition, text extraction, and multilingual text-to-speech functionalities.
2) Enable users to upload or capture images and extract textual content.
3) Support multiple languages (e.g., Hindi, English, Tamil, French) for both text translation and speech synthesis.
4) Integrate OpenTTS technology to provide natural-sounding speech output in the user's preferred language.
5) Offer text translation capabilities, allowing users to download extracted and translated text in document formats (.docx or .doc) for offline use.

C. System Implementation

The system implementation of the Android application follows a structured pipeline that integrates multiple components for seamless processing of images, text extraction, translation, text-to-speech conversion, and document generation. The architecture is designed to ensure accuracy, efficiency, and user accessibility while leveraging server-based processing for optimized performance.

The following steps outline the complete system implementation:

1) User Interaction & Image Upload

The process begins with the user interacting with the Android application. The user can either:

- Capture an image using the device camera.
- Select an existing image from the phone's gallery.

Once an image is selected, it is forwarded to the backend for processing. The application ensures a smooth user interface (UI) by incorporating a progress indicator while processing takes place.

2) Image Processing and Text Extraction

The uploaded image is passed to the Image Processing Model which identifies and extracts textual content from the image. This model is responsible for analyzing the image and detecting text regions. The extracted text is formatted into a readable structure before moving to the next stage.

3) Translation Processing

After extracting text, the Translation Model translates the content into the user's preferred language. The system supports multiple languages, such as English, Hindi, Tamil, and French, ensuring wide accessibility. The translation module ensures that the extracted content is meaningful, contextually accurate, and retains the original intent.

4) Text-to-Speech Conversion

The Text-to-Speech Converter takes the translated text and converts it into spoken output, making it accessible for users who prefer auditory content, including visually impaired individuals. The TTS system supports multiple languages, allowing users to listen to the translated text in their desired language.

5) Document File Generation & Downloading

The system also provides users with an option to download the extracted and translated text as a document file (.doc or .docx). This feature is particularly useful for users who need to save the content for future reference, academic purposes, or professional documentation. The document is then made available for download within the application.

6) User Accessibility & Output Delivery

Once the text-to-speech output is generated and the document file is ready, the user can:

- Listen to the translated content using the speech output.
- Download the document file for offline use.

II. LITERATURE REVIEW

In the paper presented by Hussain Rangoonwala1, Vishal Kaushik2, and P Mohit3, Dhanalakshmi Samiappan proposes a method for creating a complete framework that allows text to be converted into speech, text files to be converted into speech, text in different languages to be converted into speech, images to be converted into text, and images to be converted into speech, all while using the programming tool MATLAB. The various approaches are employed and then combined in an application for ease of use and accessibility.

The paper presented by K.S Bae, K.K Kim, Y.G. Chung, and W.P. Yu describes the camera-based character recognition system for mobile devices with color cameras, such as cell phones. To begin, they created a computer-based camera-based recognition system that uses techniques like image enhancement and blob coloring to extract character regions and remove noise from camera-captured images.

In the paper presented by Dr. Eleni Efthimiou, Research Director, ILSP/ATHENA R.C. states the web has developed it has become a place where people interact. They post suggestions, modify and improve each other's contributions, and share information. Dicta-Sign finds ways to enable communication between Deaf individuals by enhancing the sign language-based human-machine interfaces.

In the paper presented by David Russi, Rebecca Schneider states the Guide to Translation Project Management is a series of written instructions designed to help organizations all over the world produce high-quality translations. While it was created as a resource for National Meteorological and Hydrological Services (NMHSs) to aid in technical and training growth, the general principles apply to any agency or organization wishing to transmit information in other languages.

The paper presented by Tira Nur Fitria states that the aim of the research is to classify different types of translation techniques and determine which one is best for journal translation. A descriptive qualitative design was used in this research. Descriptive qualitative analysis provides a summary of a situation, case, or occurrence, and this approach is used to gather basic data.

In the paper presented by S. Mohideen Pillai, Dr. S. Kother Mohideen2 describes that people all over the world are becoming more interested in keeping track of their weight, eating healthy, and preventing obesity. A device that can quantify nutrition and calories in day-to-day meals is very useful, and image processing is used in such systems to accurately identify food products. For food recognition, nutrient detection, and calorie calculation, image processing techniques such as image segmentation, feature extraction, object recognition, and classification are used.

In the paper presented by U. Karthikeyan, Dr. M. Vanitha states that text recognition is a technique for extracting text from a document in the desired format (such as.doc or.txt). Pre-processing, segmentation, feature extraction, and classification are all stages in the text recognition process. The pre-processing is done to improve image enhancement while also lowering the noise signal in the input image signal. The segmentation process is used to segment the image provided online as well as each segmentation line character. In the paper presented by R. Ravi Kumar, Dr. V. Arulmozhi describes that in image processing applications image is one of the most important sources. Image processing will change human-machine interaction greatly in the future. A large variety of image processing applications and techniques helps to extract various complex features from the image. While nowadays image processing works so efficiently such that we can see what is actually present in the image. Image processing is the real core of various techniques for image enhancement. This paper discusses the overview of image processing applications, tools, and techniques. In the paper presented by Jonathan Raiman and John Miller states that the paper shows Deep Voice 3, a neural text-to-speech (TTS) system that is completely convolutional and attention-based. Deep Voice 3 is an order of magnitude faster than current neural speech synthesis systems when it comes to naturalness. Deep Voice 3 will scale up to dataset sizes never seen before in TTS. About two thousand speakers contributed over 800 hours of audio preparation. Moreover, it exemplifies.

In the paper presented by Satya Gorti and Jeremy Ma state that they compare their approach to other methods in depth for better perspective of the method. It addresses this problem by employing a captioning network to caption-generated images and exploiting the gap between ground truth and generated captions to further enhance the network

The paper presented by Jiguo Li, Xinfeng Zhang, Jia, Jizheng Xu, Li Zhang, Yue Wang, Siwei Ma, and Wen Gao states that it tries to convert speech signals to picture signals without going through the transcription stage. A speech encoder is specifically designed to represent input speech signals as an embedding function, and it is trained with a pre-trained image encoder using teacher-student learning to improve generalization capacity on new students. The paper presented by Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, Joao Carreira, Phil Blunsom, and Andrew Zisserman describes Its aim as to enhance unsupervised word mapping between languages through the use of visual grounding. The central concept is to learn embeddings from unpaired instructional videos narrated in the native language in order to create a shared visual representation of two languages[5].

III. PROPOSED METHODOLOGY

1) Research & Requirement Analysis

- Identify user needs, accessibility challenges, and potential solutions.
- Translation, LLMs, and TTS technologies to select the best approach.

2) Design & Development

- Develop the Android app UI/UX for seamless user interaction.
- Implement backend processing, integrating LLM-based, translation, and TTS functionalities.

3) Testing & Optimization

- Conduct rigorous testing with real-world images to evaluate accuracy and efficiency.
- Optimize response time for text extraction, translation, and speech output.
- Ensure proper language support and pronunciation accuracy in TTS.

4) Deployment & User Feedback

- Deploy the app on private platforms.
- Gather user feedback to refine and enhance the application's performance.

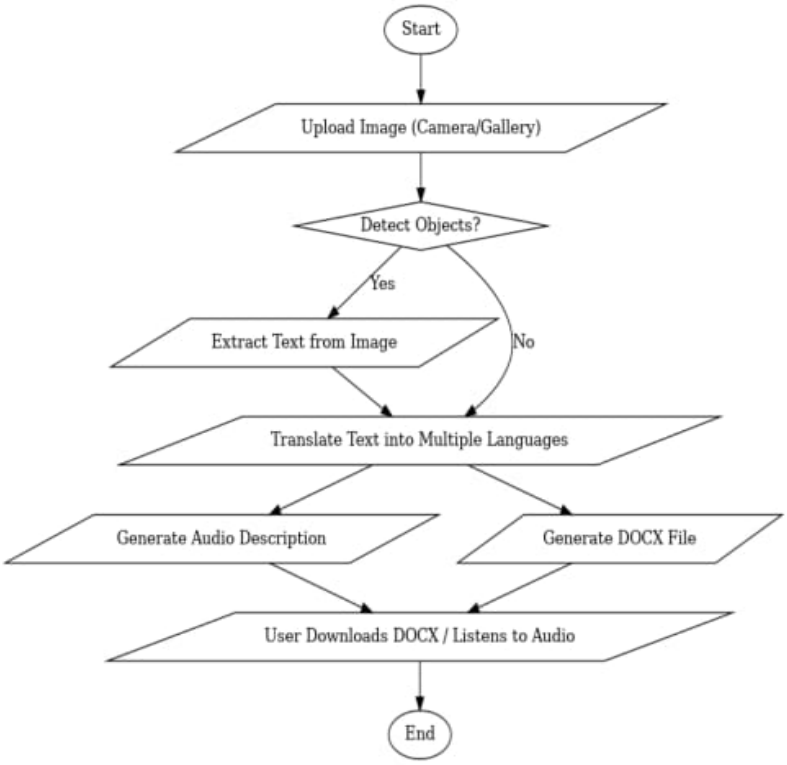


Fig 1: Image detection and Language Translation workflow

A. Application

1) Education: Converts printed materials into translated text and speech for students.
2) Accessibility: Assists visually impaired users by reading text from images aloud.
3) Travel & Tourism: Translates and vocalizes foreign-language signs and documents.
4) Business & Professional Use: Helps in translating and processing official documents.
5) Public Services: Enhances accessibility in government offices and hospitals.
6) Research & Archiving: Digitizes and translates historical or printed materials.

B. Advantages

1) Multilingual Support: Enables text extraction, translation, and speech synthesis in multiple languages.
2) Enhanced Accessibility: Assists visually impaired users by converting text into speech.
3) User-Friendly Interface: Simple and intuitive design for easy interaction.
4) Offline Document Access: Allows users to download and save translated text for future use.
5) Real-Time Processing: Provides quick and efficient text extraction, translation, and speech conversion.
6) Versatile Applications: Useful in education, business, travel, accessibility, and research.
7) Improved Communication: Bridges language barriers by converting image-based text into readable and audible formats.

C. Expected Outcome

The proposed application is expected to provide a seamless and efficient way to extract text from images, translate it into multiple languages, and convert it into speech for enhanced accessibility. Users will be able to interact with the app effortlessly, upload images, receive accurate text extraction, and access both translated text and audio outputs. Additionally, the ability to download translated content as document files will ensure offline accessibility, making the application a valuable tool for education, accessibility, and global communication.

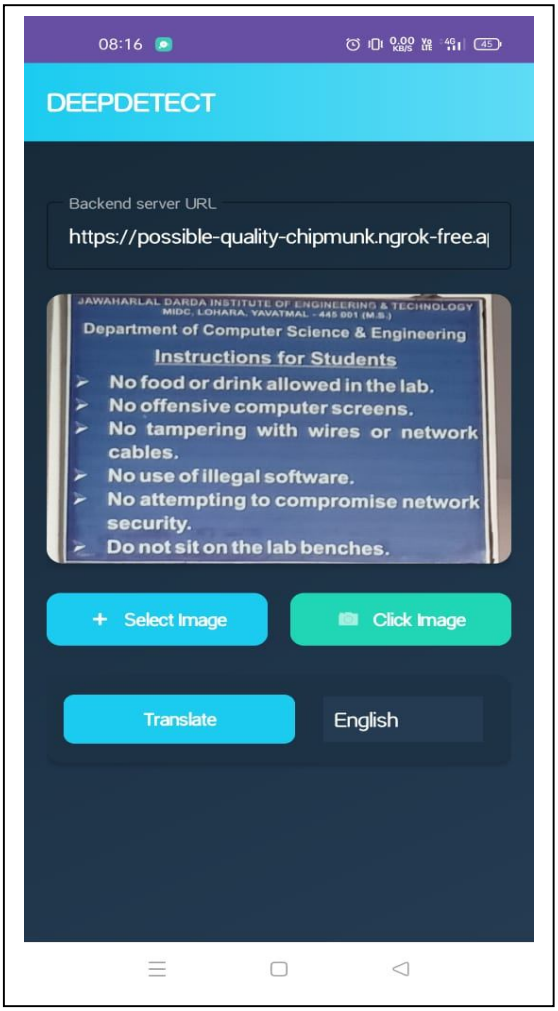


Fig 1: Image Uploads

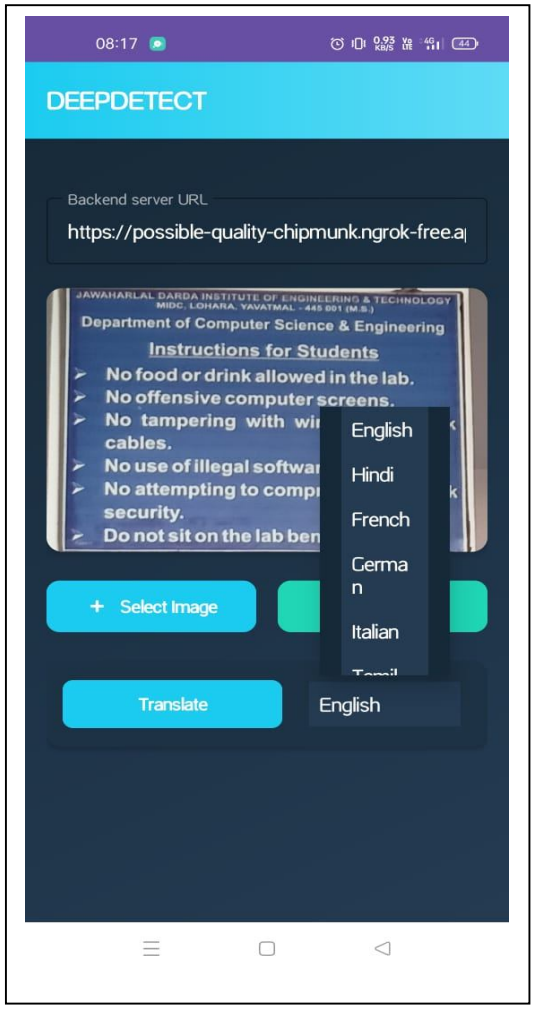


Fig 3: Select Language

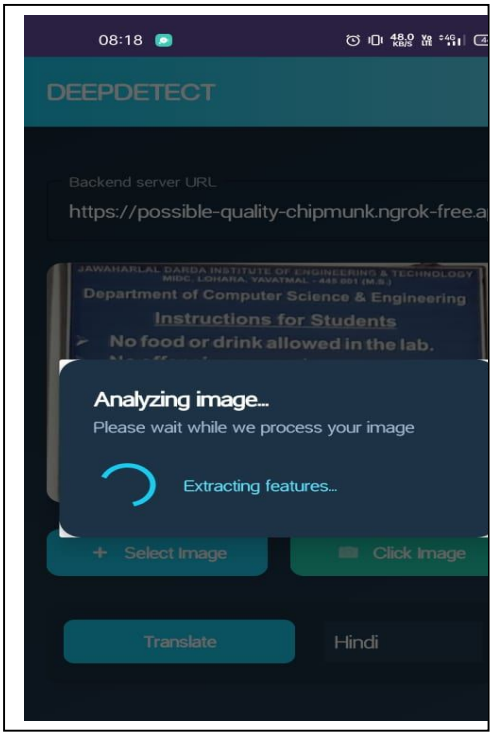


Fig 4: Analyzing Image

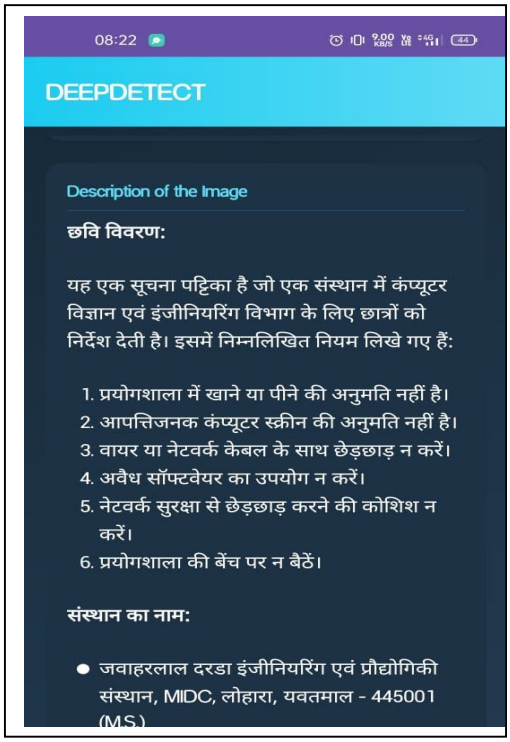


Fig 5: Download DOCX File

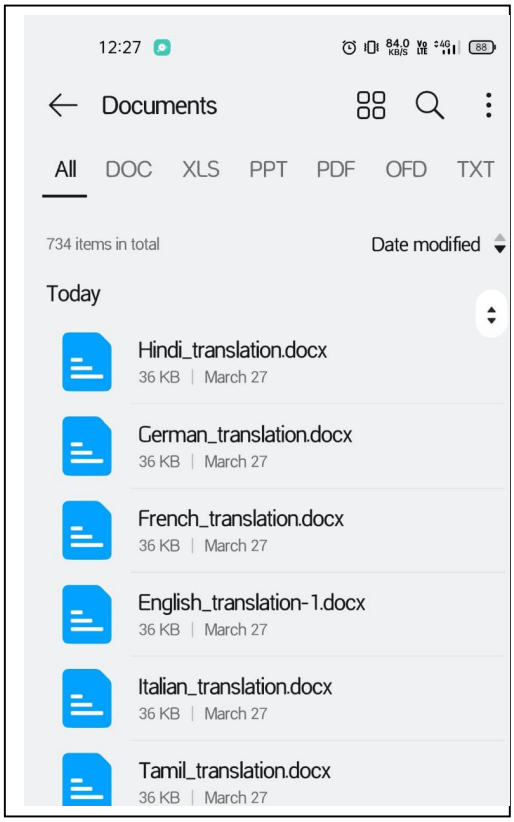


Fig 6: DOCX File in media

IV. CONCLUSION

The proposed Android application successfully integrates image processing, multilingual translation, and text-to-speech conversion to enhance accessibility and communication. By enabling users to extract text from images, translate it into multiple languages, and generate speech output, the system serves as a valuable tool for education, business, accessibility, and daily interactions. Its ability to provide downloadable document files further increases usability and convenience. With a user-friendly interface and real-time processing capabilities, this application promotes inclusivity, making information more accessible to visually impaired individuals, language learners, and global users. Ultimately, this project contributes to breaking language barriers and improving digital accessibility for diverse communities.

V. FUTURE SCOPE

The future scope of the DEEPDETECT application holds great promise as it evolves to meet the growing demands for accessibility, real-time image analysis, and multi-language support. One of the key areas for expansion is the enhancement of its language support, where the application can integrate additional languages to cater to an even broader user base, especially in regions with less representation. By incorporating more languages, DEEPDETECT can become a truly global tool for both individuals and businesses who need accurate translations and descriptions in various languages.

Additionally, the app could benefit from offline functionality, enabling users to process images, generate descriptions, and listen to translations without requiring an internet connection. This would make the app more accessible in areas with limited connectivity and reduce reliance on cloud services. Integrating edge computing or offline processing models on the device could significantly enhance the user experience, providing faster and more efficient operations even when network access is unavailable.

REFERENCES

[1] Chaitra Naik1, Amruta Khot1, Arti Jha1, Sejal D'mello2 " Text Recognition, Object Detection and Language Translation App" International Research Journal of Engineering and Technology (IRJET) Volume: 09 Issue: 03

[2] Muhammad Ajmal, Farooq Ahmad "Image to Multilingual Text Conversion for Literacy Education" 2018 17th IEEE International Conference on Machine Learning and Applications

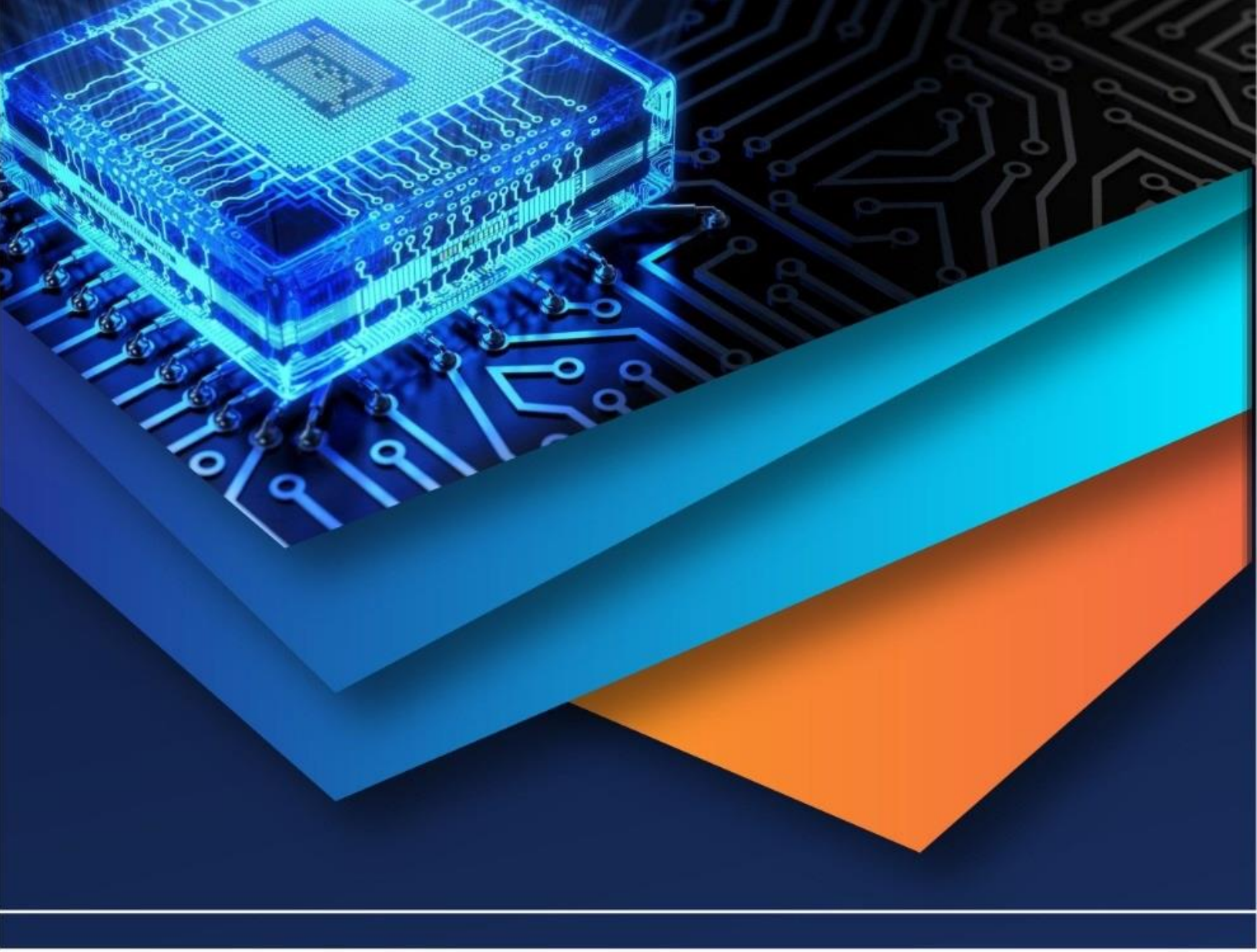
[3] Amar G. Waghade1, Anuja V. Zopate2, Ankita G. Titare3, Suraj A. Shelke4 "Text Extraction from Text Based Image Using Android" International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 03 | Mar 2018

[4] Mrs. Vishakha Shelkea, Mr. Rajat Dungarwalb, Mr. Vyom Makwanac, Mr. Keyur Babariyad "Thing Translator: An Efficient Way to Classify Different Things" International Conference on Smart Data Intelligence (ICSMDI 2021)

[5] Ojas Kumar Barawal1, and Dr Yojna Arora2 "Text Extraction from Image" International Journal of Innovative Research in Engineering & Management (IJIREM) ISSN: 2350-0557, Volume-9, Issue-3, June 2022 M. Gupta, R. R. Dhamija, R. Dias, R. V. Bidwe, G. Deshmukh, N. Jain, and S. Mishra, "Travel With Generator AI: A Novel Approach to Itinerary Creation," in 2024 IEEE Xplore. [Online]. Available: https: //ieeexplore.ieee.org/document/10775161.

[6] K. Elissa, Hiral Modi, M.C.parikh, "A Review On Optical Character Recognition Techniques", International Journal of Computer Application,2017.

[7] Huang, Rachel, Jonathan Pedoeem, and Cuixian Chen. "YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers." 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY