Object Recognition and Verbalization Tool for Early Childhood Education

Dipin Raj
Apex Institute of Technology
(CSE)
Chandigarh University
Punjab, India
dipinr505@gmail.com

Jeevan A J
Apex Institute of Technology
(CSE)
Chandigarh University
Punjab, India
jeevanaj2003@gmail.com

Rashaz Rafeeque
Apex Institute of Technology
(CSE)
Chandigarh University
Punjab, India
rashazrafeeque@gmail.com

Rhishitha T S
Apex Institute of Technology
(CSE)
Chandigarh University
Punjab, India
rhishithats002@gmail.com

Kirti
Apex Institute of Technology
(CSE)
Chandigarh University
Punjab, India
Kirtisharma230819@gmail.com

Abstract— The nature of teaching and learning process in early childhood learning makes it difficult to address the need, developmental age, ability, and interest of the learners. For early learning, this work proposes the Object Recognition and Verbalization System that is grounded on merging the Real-time Object Recognition AI and verbal cues. The proposed system allows the toddlers to explore environment, recognize the presence of the objects, and, at the same time, listen to descriptions connected with given items, which not only instills in the toddler the names of things that might interest him or her but also lets being aware of the surroundings. From the CNNs for object recognition and the NLP for speech synthesis, the tool is therefore individually operated according to each child's learning schedule and modality. Unlike a game approach, this strategy presupposes children's direct engagement in learning while presenting the essentials of understanding the surrounding world in a rational sequence. Enriching the exploration of knowledge in children, it opens vision and speech to present knowledge teaching methods in early childhood education to make it more natural and advanced.

Keywords*— *Natural Language Processing, Object Recognition, Speech Recognition, YOLO, Faster RCNN, GTTS.

I. INTRODUCTION

Childcare centers are concentrated on the first years of the child's life because it is during this period that the child forms his or her cognitive functions and speech. Although formats used in teaching and learning are helpful, they provide limited opportunity or option for variation depending on the learning gains of every child and the developmental characteristics of early children. To avoid this limitation tremendous progress is made in the areas of artificial intelligence and machine learning opened up new horizons of educational technologies that facilitate the personalized learning process. The main objective of the Object Recognition and Verbalization Tool for Early Childhood Education is to develop a learning tool that will encourage the improvement of vocabulary comprehensibility in young children and to focus their attention on secondary stimuli within their environment in real time interference.

At the heart of this tool is YOLO (You Only Look Once), which is a rapidly working object detection algorithm that can scan a child's environment and even define several objects in a few seconds. Complemented with speech recognition and text-to-speech (TTS), it immediately states the identified objects, making it easier for children to link visual and prompt auditory information. Incorporating aspects of IoT, the tool adds tangible objects into the process, which enables children to better understand objects in their day-to-day lives.

In addition to simple visual identification and vocabulary development, this project involved the development of an innovative, AI-based voice control solution whose purpose is to engage children and improve their learning experience. It is essential to point out that through the use of API technology, the tool combines different system components and can work incrementally to suit the child's learning preferences and pace. This tool is based on cognitive learning theory and employs discovery/construction/learning methodology that makes knowledge construction a constructive process and an active teaching and learning method. With these advanced technologies, the Object Recognition and Verbalization Tool is a new-world breakthrough in early childhood education to make education fun, individually tailored, and productive.

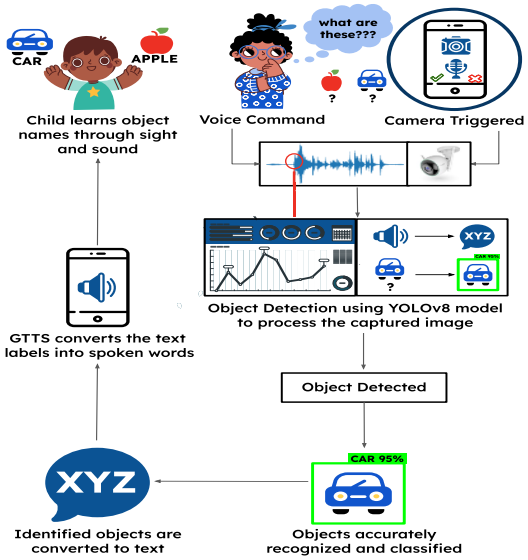

Figure 1: End-to-end process from speech input to spoken output

II. LITERATURE SURVEY

1. YOLO:

YOLO – You Only Look Once – is an advanced object detection techniques that is faster and more accurate than other comparable methods making it ideal for real-time applications in this project. Since it can identify a number of objects on a single frame with reasonable time consumption, the system can respond immediately, which is critical to maintain young children's attention. In early childhood education, YOLO enables the swift identification of many objects in a child's environment while keeping the system quick and engaging. Such capability enables the tool to be responsive to new environment and learning environments for individual and contextualized learning.

2. OBJECT RECOGNITION:

Real-time object recognition is a core aspect of this project because it helps the designed system to recognize and categorize objects in a child's environment. The system, driven by CNNs, recognizes familiar and ubiquitous objects thus enabling the system to produce verbal descriptions of the objects. This feature is especially useful in teaching young children as it functions as a word bank, that introduces new words in relation to familiar objects alongside refining a child's comprehension of the world around them. Real-time object recognition is the key to the interaction between the child and the tool, as it returns an immediate reaction in terms of visual and auditory feedback which make the learning process more engaging.

3. SPEECH RECOGNITION/ TEXT-TO-SPEECH (TTS):

The integration of speech recognition and text-to-speech (TTS) technology allows the system interact to the child, making learning process more engaging with using natural language. When speech recognition is applied it allows the system to answer simple voice commands/ questions from the children hence making it an interactive system in teaching. TTS turns text descriptions of recognized objects into audio feedback making the link between visual recognition and auditory learning. Furthermore, TTS makes this learning approach even more unique by varying the speech rate according to the development level of individual children thus creating an even more effective learning model.

4. IOT TECHNOLOGY:

The IoT enhances this project significantly and make it possible to connect devices and sensors to shape smart learning environment. IoT makes it easy to link physical objects with digital feedback, and it also enables the system to identify objects not only with help of vision but with the help of touch or proximity sense. Using IoT connected devices the system can recognize and provide information about objects around the environment immediately, providing the learner with context and richness of the environment to create learning interactions that are curiosity driven and explorative in nature.

5. VOICE ASSISTANT LEARNING:

Voice assistants, powered by artificial intelligence enhance this proposal through providing a natural and conversational layer between the child and the system. These assistants are capable of offering customized learning experiences for each child by responding to the unique question that the kid may ask, and the progress in learning that the child may have made. Making use of voice-based interactions, the system provides children with the best opportunity to engage in activities that are related to education and would contribute to their cognitive development and would improve their language skills. The use of voice assistant technology alongside real-time object recognition maintains an effective and contextually relevant learning process which is also persistent.

6. COGNITIVE LEARNING:

Cognitive learning theory emphasizes learners as active creators of knowledge and, therefore, this project adopts the approach of providing an engaging as well as an explorative learning environment. Using real time object recognition and verbal prompts, the tool trains children to map between visual and auditory inputs and therefore enhances language and concept development. Cognitive development is supported by Interactive challenges and prompts which in return promotes problem-solving and critical thinking. This approach aligns with the principles of cognitive learning theory as it invites the children to learn on their own, to query whatever situation they are in, and to obtain pertinent and timely feedback that consolidates the content of their minds and enhances their processing abilities.

7. API TECHNOLOGY:

This project relays on Application Programming Interfaces (APIs) which help to integrate primary components – object recognition, TTS, voice assistant. Real-time and interactive operational methods are characterized by APIs that enable general communication between modules. For example, TTS in Google or language models in OpenAI may recite object descriptions making a child's learning process much more enjoyable. This gives the system a modular design which can easily allow for future integration of new tools and improvements. By upholding the adaptive, responsive nature of the tool, APIs hence ensures that it can grow alongside a child's learning journey.

8. COMPUTER-ASSISTED EDUCATION:

Computer-assisted education is at the heart of this project, creating a link between digital tools and the familiar ways young children learn. By bringing AI and machine learning into the mix, this project introduces a fresh approach to learning that adds interactive technology to traditional teaching. The tools—like the object recognition and verbalization system—make learning personal, letting each child learn in the way that fits them best. Children hence can explore and get immediate feedback making learning at their own pace through fun and hands-on experiences. This approach transforms early education, creating a journey that's truly built around them with technology adapting to each child's unique style and rhythm.

III. PROPOSED SYSTEM

The proposed system aims to dynamically creates an educational tool that leverages object recognition and speech synthesis to foster interactive learning for young children which is powered by AI. At its core, the system uses computer vision to detect and identify objects in real time, responding with spoken feedback to help children connect visual information with auditory cues. It's like having a learning companion that helps children discover the world around them, making vocabulary building as simple and enjoyable as pointing, asking, and learning. Through translating simple everyday activities in the environment into learning processes the children delightfully get engaged and develop language ability simultaneously.

The system is designed to create a smooth, adaptable learning experience by combining components that work together effortlessly. By simply using voice commands, kids can activate object recognition, and the technology will recognize what's in front of them, describing it back in easy-to-understand terms. In case of successful detection, the system produces verbal descriptions of the objects and instant auditory feedback is produced. This real-time interaction is conscious, and aimed at entraining children

actively, and giving them immediate positive feedback which is timely and catches up with the pace at which children learn naturally. Being a flexible tool, the vocabulary, images dataset and response styles can easily be modified in order to fit each age range as required. It progresses with the child and also with the interest that the child has at that particular time; thus, making learning to be more of fun as it follows the phases of the child's discovery.

Furthermore, the system is also friendly and flexible, capable of being accessed using any form of device be it a personal computer, a mobile phone and more. This way, while being compatible with standard hardware and utilizing simple software, the proposed system can reach a wide audience including educators, caregivers, and young learners at home. The features of this application include the ability to learn according to a number of methods and is adaptable to the child's interest, which makes learning more fun for the child. Hence it has the potential to create a positive shift in early childhood education with the help of Artificial Intelligence to engage Children in learning meaningfully.

IV. METHODOLOGY

The real-time object detection and verbalization system that has been developed incorporates computer vision and speech recognition. The main goal was to develop a tool for object recognition through a webcam connection initiated through a voice command, and with results communicated through a speech output. The following illustrates more of how the system implementation will be done, the set up as well as the workflow.

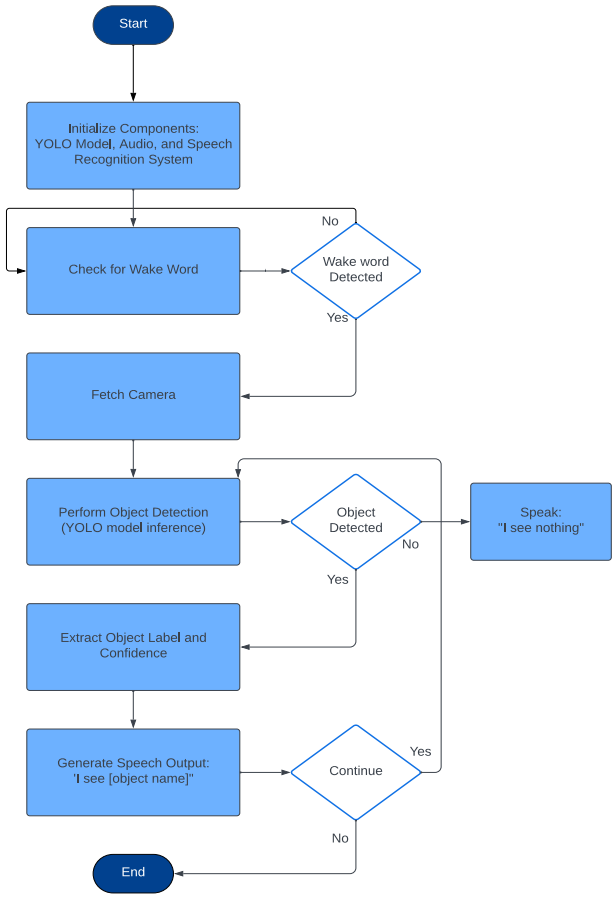

Figure 2: Visual of object recognition and speech feedback flow

Computer vision and speech recognition are combined to design the real-time object detection and verbalization system, which allows recognizing objects through a webcam and reacting with speech. The main goal is to enable a user to perform real-time object recognition by starting a video stream using a particular voice command, with the outcome being provided in the form of audio feedback. The system setup involves two main hardware components: recalling voice commands in relation to an object and using a microphone and a camera to interact with it. This setup can either use a locally connected web cam or a mobile phone camera connected through the network or any other live network cam which is preferably hardware programmable. The software environment of the system includes Python application with the use of Open CV for real time image capturing and processing, Torch for neural network management, speech_recognition for voice recognition and gTTS for giving out output voice. As real-time response is a major aspect, inference time is prioritized over its quality, and for object detection, the current state-of-the-art model YOLOv8 is used. Using the YOLO algorithm, fine-tuned on the COCO dataset, we can quickly detect a vast number of ordinary objects.

The entire workflow is divided into three key phases: speech recognition of the voice commands, identification of the objects within the environment, and generation of speech feedback. At the start of the process, the system listens and expects specific verbal prompts like "What is this?". The microphone records voice inputs which are translated by the speech recognition library to activate the system. After the detection of the voice command, the object detection pipeline is initiated. If this command is recognized by the module, the system turns on the video stream to start object recognition. The camera feed is started by OpenCV from a networked video stream or a local camera. The stream is then scaled properly for real-time data processing with YOLOv8. After that, YOLOv8 characterizes the objects and classifies them into coordinates and categories, respectively by going through each frame. The bounding boxes that enclose the real objects serve to give faster confirmation of what the detections are inside the video feed.

After the object detection, the scene is described through sounds via auditory output. For example, if a cat and a book are detected, it forms the statement "I see a cat and a book." The descriptive text generated is then taken by gTTS to generate speech. The audio output is in the pygame library which makes the system speak what it sees. This verbal feedback, given in finite state messages, consists of a user-friendly interface that is easily able to communicate detected objects to the user. By designing such an end-to-end process, the system integrates both the vision and speech components to support the interaction and in-sequence object identification and verbalization for real-time supportive conversation.

V. RESULT

Real-time object identification and verbalization were successfully shown by the built model, which was able to recognize items in a variety of real-world scenarios. The system correctly identified a variety of things in the surroundings, including everyday objects like books, plants, and toys, by utilizing a webcam or mobile device camera. This aligning with the intended vocabulary set for early learners. The YOLOv8 model proved stable in the simultaneous detection of multiple objects and provided immediate visual feedback necessary to sustain the child's captivation.

The model's performance was assessed throughout the testing phase using a variety of objects, including both well-known ones (like "Apple," "Banana," and "Cup") and more complicated ones (such "Elephant" and "Teddy Bear"). For each object, preprocessing and inference times, along with the confidence score, were recorded. The system's overall accuracy and efficiency can be inferred from these findings, which are averaged across five test runs. Preprocessing times were comparatively constant across

items, averaging between 3.0 and 6.0ms. In contrast, inference times varied more dramatically, ranging from 81.5ms for a "Book" to 145.4ms for a "Cup." The model showed great accuracy in spite of this unpredictability, with confidence ratings continuously over 0.80 and as high as 0.96 for "Car." This consistency in detection confidence indicates the system's robustness, even when identifying diverse objects in real time. The technology is appropriate for an interactive, kid-focused learning environment where instant feedback is needed because the performance metrics demonstrate its capacity to analyze and transmit results quickly.

Figure 3: System capturing and identifying objects accurately

The system's capacity to precisely capture and identify objects is seen in Figures 4, 5, and 6, which also prominently display labels and bounding boxes surrounding identified items. This gives young users an easy-to-understand visual confirmation. High responsiveness was attained by the model, which processed each frame quickly enough to allow for fluid, real-time interaction. Bounding boxes and object labels quickly arose around identified objects, which strengthens interaction through unambiguous visual feedback. An immersive and conversational learning experience was supported by the smooth interaction made possible by the continuous, real-time feedback loop, which allowed things to be detected and verbally described almost instantly.

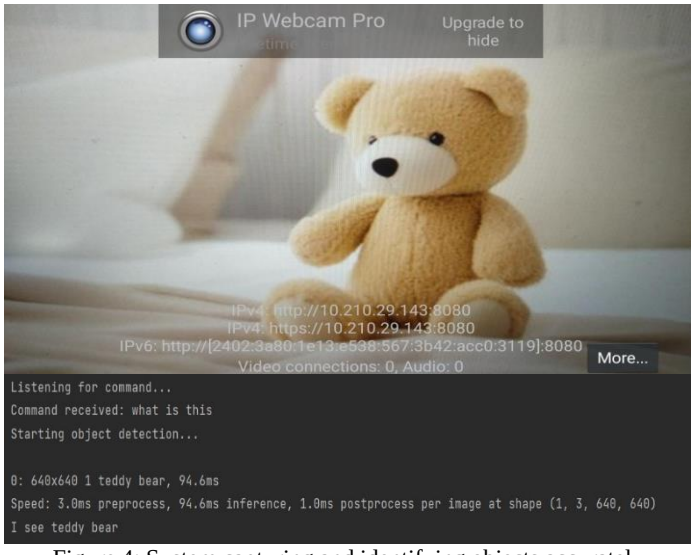

Figure 4: System capturing and identifying objects accurately

Table 1: Preprocessing and inference times during system tests

Object	Preprocess Time	Inference Time	Confidence
Apple	4.9ms	100.7ms	0.89
Banana	5.0ms	96.7ms	0.82
Book	4.5ms	81.5ms	0.81
Car	4.0ms	96.1ms	0.96
Cup	4.5ms	145.4ms	0.93
Elephant	6.0ms	129.3ms	0.92
Scissors	4.5ms	130.9ms	0.93
Teddy Bear	3.0ms	94.6ms	0.89

Figure 5: System capturing and identifying objects accurately

Clear and concise descriptions, such "I see a book and a toy," were supplied via the speech output, which was produced using gTTS and sent via the pygame audio module. This real-time verbalization not only enabled the child to link the words being spoken with corresponding cues in the child's environment but also made it easier to teach new vocabulary. The system's efficacy as a tool for early childhood education was confirmed by the combination of rapid recognition, precise labeling, and instantaneous aural feedback. It also revealed a possibility of enhancing learning activities by making them interesting, unique, and communicative.

VI. CONCLUSION

Improving early learning via the Object Recognition and Verbalization Tool is a potent illustration of how cutting-edge technology may improve early childhood education. This program utilizes artificial intelligence and machine learning particularly the YOLO object identification model to make an innovative learning platform that matches a child's cognitive development level with engaging activities that teach a customized vocabulary set. Along with text-to-speech and speech recognition features, the system provides real-time visual feedback when kids engage with it. Through active participation and the tool's verbal explanations of the items they see, children are able to make the connection between images and meaning. This helps broaden their understanding and also expand their perceptive lexicon. The program deepens their learning by encouraging them to physically interact with real-world items which goes beyond just naming objects. It helps them build a stronger, more intuitive understanding of the objects around them which reinforces what they hear using hands-on experience.

Additionally, the AI-based voice assistant enhances learning by provoking or eliciting learner interest and engagement in young learners. Exploratory and problem-solving methods based on cognitive learning theories are supported by API technology, which guarantees that the tool can adjust to each child's particular learning style and speed. The technology actively engages youngsters by customizing the educational experience, which raises interest and improves learning results. This is an important direction of the development of AI in educational technologies, and such resources provide important steps toward making early childhood education more engaging, individualized, and effective in preparing children for further learning.

REFERENCE

[1] Hanafi, H.F., Wong, K.T., Adnan, M.H.M., Selamat, A.Z., Zainuddin, N.A. and Lee Abdullah, M.F.N., 2021. Utilizing Animal Characters of a Mobile Augmented Reality (AR) Reading Kit to Improve Preschoolers' Reading Skills, Motivation, and Self-Learning: An Initial Study. *International Journal of Interactive Mobile Technologies*, *15*(24).

[2] Wu, Q., Wang, S., Cao, J., He, B., Yu, C. and Zheng, J., 2019. Object recognition-based second language learning educational robot system for Chinese preschool children. *IEEE Access*, *7*, pp.7301-7312.

[3] Qi, S., Ning, X., Yang, G., Zhang, L., Long, P., Cai, W. and Li, W., 2021. Review of multi-view 3D object recognition methods based on deep learning. *Displays*, *69*, p.102053.

[4] Bazargani, J.S., Sadeghi-Niaraki, A., Rahimi, F., Abuhmed, T. and Choi, S.M., 2022. An iot-based approach for learning geometric shapes in early childhood. *IEEE Access*, *10*, pp.130632-130641.

[5] Rahiem, M.D., 2021. Storytelling in early childhood education: Time to go digital. *International Journal of Child Care and Education Policy*, *15*(1), p.4.

[6] Zaini, N.A., Noor, S.F.M. and Wook, T.S.M.T., 2019. Evaluation of api interface design by applying cognitive walkthrough. *International Journal of Advanced Computer Science and Applications*, *10*(2).

[7] Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen W. White. 2019. VERSE: Bridging Screen Readers and Voice Assistants for Enhanced Eyes-Free Web Search. In Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19). Association for Computing Machinery, New York, NY, USA, 414–426.

[8] E. de la Guía, V. L. Camacho, L. Orozco-Barbosa, V. M. Brea Luján, V. M. R. Penichet and M. Lozano Pérez, "Introducing IoT and Wearable Technologies into Task-Based Language Learning for Young Children," in IEEE Transactions on Learning Technologies, vol. 9, no. 4, pp. 366-378, 1 Oct.-Dec. 2016, doi: 10.1109/TLT.2016.2557333.

[9] Mevlüde Akdeniz, Fatih Özdinç, Maya: An artificial intelligence based smart toy for pre-school children, International Journal of Child-Computer Interaction, Volume 29, 2021, 100347, ISSN 2212-8689.

[10] Khondaker A. Mamun, Rahad Arman Nabid, Shehan Irteza Pranto, Saniyat Mushrat Lamim, Mohammad Masudur Rahman, Nabeel Mahammed, Mohammad Nurul Huda, Farhana Sarker, Rubaiya Rahtin Khan, Smart reception: An artificial intelligence driven bangla language-based receptionist system employing speech, speaker, and face recognition for automating reception services, Engineering Applications of Artificial Intelligence, Volume 136, Part A, 2024, 108923, ISSN 0952-1976

[11] Amara, K., Boudjemila, C., Zenati, N., Djekoune, O., Aklil, D., & Kenoui, M. (2022). AR Computer-Assisted Learning for Children with ASD based on Hand Gesture and Voice Interaction. *IETE Journal of Research*, *69*(12), 8659–8675.

[12] : **arXiv:2401.05459** (cs) [Submitted on 10 Jan 2024 (v!), last revised 8 May 2024 (this version, v2)] Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security

[13] Devi, J.S., Sreedhar, M.B., Arulprakash, P., Kazi, K. and Radhakrishnan, R., 2022. A path towards child-centric Artificial Intelligence based Education. *International Journal of Early Childhood*, *14*(3), pp.9915-9922.

[14] Fitria, T.N., 2021, December. Artificial intelligence (AI) in education: Using AI tools for teaching and learning process. In *Prosiding Seminar Nasional & Call for Paper STIE AAS* (Vol. 4, No. 1, pp. 134-147).

[15] Ganesh, D., Kumar, M.S., Reddy, P.V., Kavitha, S. and Murthy, D.S., 2022. Implementation of AI Pop Bots and its allied Applications for Designing Efficient Curriculum in Early Childhood Education. *International Journal of Early Childhood Special Education*, *14*(3).

[16] Alam, A., 2022, April. A digital game based learning approach for effective curriculum transaction for teaching-learning of artificial intelligence and machine learning. In *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)* (pp. 69-74). IEEE.

[17] Ng, D.T.K., Lee, M., Tan, R.J.Y., Hu, X., Downie, J.S., & Chu, S.K.W. (2023). A review of AI teaching and learning from 2000 to 2020. Education and Information Technologies, 28(7), 8445-8501. https://doi.org/10.1007/s10639-022-11312-3

[18] Lin, S.Y., Chien, S.Y., Hsiao, C.L., Hsia, C.H. and Chao, K.M., 2020. Enhancing computational thinking capability of preschool children by game-based smart toys. *Electronic Commerce Research and Applications*, *44*, p.101011.

[19] Qureshi, K.N., Kaiwartya, O., Jeon, G. and Piccialli, F., 2022. Neurocomputing for internet of things: object recognition and detection strategy. Neurocomputing, 485, pp.263-273.

[20] Adarsh, P., Rathi, P. and Kumar, M., 2020, March. YOLO v3-Tiny: Object Detection and Recognition using one stage improved model. In 2020 6th international conference on advanced computing and communication systems (ICACCS) (pp. 687-694). IEEE.

[21] Amit, Y., Felzenszwalb, P. and Girshick, R., 2021. Object detection. In Computer Vision: A Reference Guide (pp. 875-883). Cham: Springer International Publishing.

[22] Rahman, M.A. and Sadi, M.S., 2021. IoT enabled automated object recognition for the visually impaired. Computer methods and

programs in biomedicine update, 1, p.100015.

[23] Hussan, M.I., Saidulu, D., Anitha, P.T., Manikandan, A. and Naresh, P., 2022. Object Detection and recognition in real time using deep learning for visually Impaired people. International Journal of Electrical and Electronics Research, 10(2), pp.80-86.

[24] Guravaiah, Koppala, Yarlagadda Sai Bhavadeesh, Peddi Shwejan, Allu Harsha Vardhan, and S. Lavanya. "Third eye: object recognition and speech generation for visually impaired." *Procedia Computer Science* 218 (2023): 1144-1155.