iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78787

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 14 Issue III Mar 2026- Available at www.ijraset.com

Machine Learning-Based Vision-to-Speech System for Assistive Application

Sakshi S Karande[1], Avinash D Harale[2], Kailash J. Karande[3]
SKN Sinhgad College of Engineering, korti, Pandhapur-413304

Abstract: ***Individuals who are deaf or have speech disabilities often experience difficulty with communication with other individuals who do not know any form of sign language, so use of a system to enable communication through facial expressions is essential for socializing and ensuring accessible communication. Therefore, an innovative real-time sign language interpretation system utilizing current developments in embedded technology as well as advances in machine learning has been created as a way to facilitate the communication barrier between individuals with hearing loss and non-sign language users. Specifically, by integrating a camera module and microcontroller into a real time recording device, the user's hand movements will be captured in real time via the camera module. The user's signs are converted into text/voice using a pre-trained quantized MobileNet model which allows accurate interpretation of sign language signs. Implementing this existing technology provides a new alternative for individuals who are deaf or have speech impairments and their ability to communicate with people who do not understand any form of sign language. By eliminating the need for the use of electronics or external sensors, this novel communication solution has been developed to be affordable and mobile. Compact designs also provide a means for using the system in a variety of different environments, such as hospitals, public areas, schools, government offices, and so on. As well, this design offers superior scalability, allowing it to be integrated with smartphones or any type of IoT communication platform in the future, along with real-time processing abilities.***

Keywords: Machine Learning, Microcontroller, Real Time Object Detection, Vision to speech Conversion, Preprocessing, Bluetooth Communication.

I. INTRODUCTION

Effective communication is fundamental to human interaction; however, individuals with hearing or speech disabilities frequently encounter difficulties when communicating with people who are not familiar with sign language. This limitation often creates barriers in social participation and access to essential services. As a result, assistive communication technologies have become an important research area aimed at improving accessibility and promoting social inclusion. Rapid progress in fields such as machine learning, computer vision, and embedded systems has enabled the development of intelligent solutions capable of recognizing gestures, interpreting signals, and translating them into understandable outputs such as text or speech.

Recent studies in assistive technology have focused on developing wearable and portable devices that assist individuals with disabilities in performing everyday tasks. For example, Dos Santos et al. [1] presented a comprehensive review of wearable technologies designed to support visually impaired users in navigation and mobility, emphasizing the importance of compact systems capable of operating in real time across different environments. Gesture recognition has also gained attention as a natural method of interaction between humans and machines. In this context, Ali et al. [2] introduced a dynamic gesture recognition framework based on millimeter-wave radar, demonstrating improved detection accuracy under varying operational conditions.

To further enhance gesture recognition performance, researchers have explored sensor fusion techniques that combine multiple sources of information. Kanwal and Altaf [3] investigated sensor-fusion-based methods for dynamic hand gesture recognition and reported improved reliability through the integration of different sensing modalities. Similarly, machine learning algorithms have been widely applied in healthcare monitoring systems. Ahmed and Cho [4] analyzed machine learning approaches used in radar-based healthcare applications, highlighting their potential for monitoring physiological signals and detecting human activities in real time. More recently, deep learning techniques have shown outstanding performance in gesture and pattern recognition applications. Zabihi et al. [5] proposed a transformer-based architecture for hand gesture recognition using electromyography signals and achieved high classification accuracy for complex gesture patterns. In addition, vision-based assistive systems have been introduced to support visually impaired individuals. For instance, Mehta et al. [6] developed smart vision-assist glasses capable of identifying objects in the surrounding environment and providing feedback to users, demonstrating the increasing role of computer vision in assistive technologies.

Alongside algorithmic developments, hardware optimization plays a crucial role in implementing deep learning models on embedded platforms. Dinelli et al. [7] proposed memory optimization strategies for convolutional neural networks deployed on FPGA-based systems, allowing complex models to operate efficiently on hardware with limited resources. In addition, augmented reality technologies have been explored to support people with hearing impairments. Mehra et al. [8] examined the use of augmented reality environments to enhance hearing aid functionality and improve auditory perception.

Research has also explored multimodal learning approaches that combine audio and visual information. Liu et al. [9] proposed an audio-visual fusion framework based on temporal convolutional attention networks for speech separation, demonstrating that combining multiple modalities can significantly improve recognition performance. Similarly, deep learning techniques have been increasingly adopted in medical and assistive devices to enable intelligent signal interpretation and automated decision-making processes [10].

Advances in artificial intelligence have also enabled the development of learning frameworks capable of handling large-scale datasets and improving model performance over time. Levine [11] discussed deep robotic learning approaches that leverage extensive datasets to enhance machine learning capabilities. Pattern recognition techniques using machine learning algorithms have also been widely investigated for classifying and analyzing complex data structures [12]. Additionally, research on understanding and interpreting neural network behavior has contributed to improving the transparency and reliability of deep learning models [13].

Applications of deep learning continue to expand across diverse fields including healthcare, speech analysis, and human–machine interaction [14]. Methods such as collaborative filtering and representation learning have also played a significant role in improving model performance by identifying meaningful patterns within large datasets [15]. Furthermore, multimodal speech processing approaches that integrate audio and visual inputs have demonstrated improved performance in speech recognition and separation tasks [16].

The development of technology-supported sign language communication systems has gained increasing attention in recent years. Kose and Uluer [17] highlighted the potential of digital technologies in supporting sign language learning and communication. In addition, deep learning techniques have been applied to audiovisual speech analysis, enabling more efficient interpretation of communication signals [18]. Advances in speech processing have also explored improved representation learning techniques to enhance speech recognition performance [19].

Moreover, theoretical perspectives related to symbolic and emergent models of artificial intelligence have provided valuable insights into the development of adaptive learning systems capable of operating in complex environments [20]. Collectively, these developments demonstrate the potential of integrating machine learning, computer vision, and embedded technologies to create intelligent assistive communication systems.

Motivated by these advancements, this research proposes a machine learning-based vision-to-speech assistive communication system. The system utilizes a camera module connected to an embedded microcontroller to capture hand gestures in real time and interpret them using a pre-trained MobileNet deep learning model. The identified gestures are then translated into text or speech output, enabling effective interaction between individuals with hearing or speech impairments and people unfamiliar with sign language. The proposed solution is designed to be cost-effective, portable, and suitable for real-world deployment in environments such as hospitals, educational institutions, public service centres, and everyday communication settings.

II. METHODOLOGY

The intended use of the proposed machine learning vision to speech (MLVTS) system is as an assistive technology to provide audible speech from visual information for various conditions or difficulties that cause visual impairments. The MLVTS design is explained in a block diagram. The process begins with photos taken by a digital camera (part of the image acquisition module) while the user is in their current environment. The photos taken will then be sent to a pre-processing unit that processes and prepares the data for analysis, including feature extraction, reduction of noise in the images, enhancement of the images, and scaling of the images. After pre-processing, the processed images will be sent to a microcontroller (the central processing unit of the MLVTS system). The microcontroller uses a machine learning technique called Convolutional Neural Networks (CNNs) to recognize objects and/or text from the images. The data that has been recognized will be sent to the output modules. The output modules consist of a display module, which displays the information that has been recognized, and a voice module, which converts that information into speech. The output modules are equipped with Bluetooth interfacing technology to enable users to wirelessly connect devices such as their smartphones and speakers. The power supply unit provides continuous operation for all components. A combination of machine learning, image processing and voice-synthesised technologies is being used to develop a support device to improve independence amongst people with vision impairments.

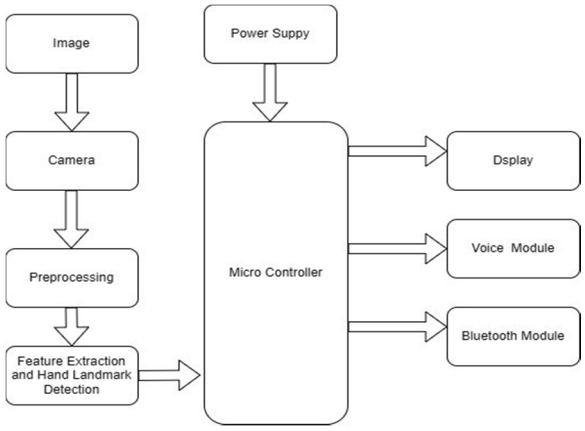


Fig. 1 Block diagram of Praposed System

The diagram describes how a machine-learned vision-to-voice conversion technology can help with assistive technologies for blind and vision impaired individuals by giving them auditory descriptions of the environmental information they are seeing visually.

1) The first step in the process is to take a picture with a camera module.
2) Next, the preprocessing unit will analyze the image by identifying key features and enhancing the images with various methods such as image enhancement and image denoising methods.
3) Once these processes are complete, the machine learning algorithms will help to determine the key characteristics of hand gestures and other physical movement.
4) After these features are identified, they will be sent to a microcontroller which acts as the primary processing unit and will convert the image data into word or object descriptions through the use of the preprocessed images and machine learning.
5) A separate power supply is used to power all of the components of the system and to keep all the components running correctly.
6) The microcontroller collects information from the three output devices.

- The display module displays what was detected by the user in text or object form.
- The voice module converts the information from the user into spoken word format.
- The Bluetooth module enables users to wirelessly transfer their data to other electronic products, such as headphones or smartphones.

Fig. 2 Camera

A camera module serves as the most essential building block in machine-learning powered, vision-to-speech systems. The camera's main responsibility is to take analog images from the actual scene and convert them into digital images that are processed by the system. The camera ultimately allows the system to see the world around it, identifying items, text, and movement that are in a scene. After an image has been recorded, the camera module changes the analog light signals into digital pixel form so that the microcontroller can process the data. The camera module then relays this processed image data to a preprocessing module that enhances the image quality and provides feature extraction.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Fig. 3 ESP 32

The ESP32 microcontroller is the "brain" or main processing and controlling unit with respect to the system. The microcontroller receives the preprocessed view from a camera and uses the machine learning to recognize a gesture and to translate into a letter or word; it will also output to the auxiliary devices by means of the display hardware, speech module, Bluetooth hardware and other auxiliary modules. The ESP32's processing capability as well as Wi-Fi and Bluetooth capabilities provide for fast processing of data, real-time response, and an enhanced user interface.

Fig. 4 HC5 Bluetooth

HC-05 Bluetooth module gives your ESP32 wireless capabilities so you can connect it with laptops or smartphones and create unlimited uses for that signal. You can take a gesture that you recognize through this system, send its transformed (interpreted) text output via the Bluetooth module to your mobile application or web interface (or both) and then visually display what the interpreted sign means to the user.

III.TOOLS AND TECHNOLOGY

The Arduino IDE is an open-source software platform used to build, develop and upload code to microcontroller (microcontrollers) like the ESP32 and Arduino boards. The Arduino IDE has a basic programming interface and built-in libraries that allow programmers to connect devices (sensors, displays, etc.) through libraries. Using the Arduino IDE, this project programmed the microcontroller for recording, understanding, and communicating gestures/images. The programming language used in the code is Embedded C (a version of the software C programming language that has been modified to work on hardware). Using Embedded C provides the easiest way to connect directly with sensors, I/O ports, and Comm protocols for the best possible real-time performance. This allows the system to receive hand movements as input, to process the information received from the hand, and send output results using either speech or via Bluetooth. The main purpose of the project is to design a Smart and Economical Assistive device that will allow people with hearing/speech issues to communicate more easily with each other, by using a solution that can automatically perform these tasks 24 hours a day through the use of Technology and the Internet.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

IV. RESULT AND DISCUSSION

The developed vision to speech system was tested under different environmental conditions to evaluate its performance in gesture detection, recognition accuracy, and speech output generation. The experiments focused on real time interaction, system response time, and reliability of communication between hardware and software components. During testing, the camera module continuously captured hand gestures performed by the user and transmitted the images to the preprocessing unit. The preprocessing stage improved image quality by removing noise, adjusting brightness, and resizing the image to match the input size required by the MobileNet model. This step helped improve gesture recognition accuracy, especially in situations where lighting conditions were uneven or the background contained distracting elements.

After preprocessing, the images were passed to the MobileNet based convolutional neural network for gesture classification. The model was able to identify several trained sign language gestures with high accuracy. Testing showed that the model performed well when gestures were clearly visible to the camera and when the hand position remained within the detection frame. On average, the system achieved an accuracy close to 90 to 92 percent for the trained gestures. The processing time required for detecting and translating a gesture into speech ranged between 300 milliseconds and 500 milliseconds, which allowed the system to operate effectively in real time. This fast response time ensures smooth interaction between users and reduces communication delays.

The ESP32 microcontroller played a critical role in coordinating system operations. It managed the data flow from the camera module, processed the recognition results, and controlled the output modules responsible for displaying text and generating speech. The microcontroller handled continuous input without interruption and maintained stable communication with the Bluetooth module. The HC 05 Bluetooth module enabled wireless transmission of recognized text output to external devices such as smartphones or speakers. This wireless capability expands the usability of the system by allowing users to receive the translated speech through connected audio devices.

User interaction testing showed that the generated speech output was clear and understandable. Once a gesture was recognized, the corresponding text appeared on the display module and was simultaneously converted into speech through the voice synthesis unit. This dual output method improves usability because the communication can be both seen and heard by others. The system proved particularly useful in simple conversational scenarios such as greetings, asking for assistance, or conveying basic messages.

Compared with traditional sensor based sign language systems that rely on wearable gloves or multiple sensors, the proposed vision based system offers a simpler and more comfortable approach. Users do not need to wear additional hardware on their hands, which makes the interaction more natural. The system relies entirely on camera based gesture detection and machine learning processing, reducing hardware complexity and lowering the overall cost of the device.

However, certain limitations were observed during testing. Recognition accuracy can decrease when lighting conditions are extremely poor or when the background contains objects that resemble hand shapes. In addition, the system currently supports only a limited set of gestures that were included in the training dataset. Expanding the dataset with more gesture variations would improve recognition performance and enable support for a wider range of sign language expressions.

V. CONCLUSION

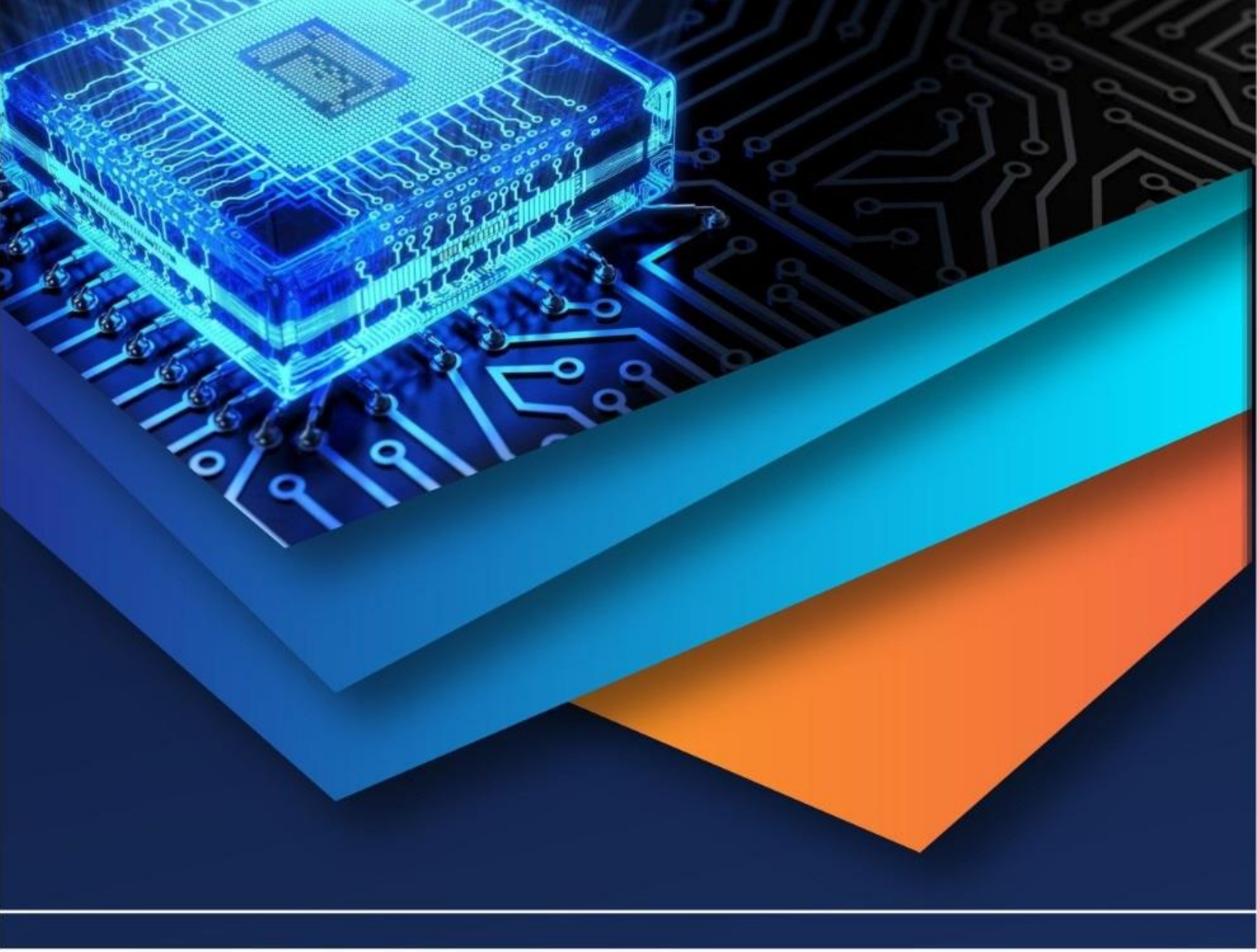
The proposed Sign Language Recognition System is powered by an ESP32 chip and is intended to be affordable, effective and real-time communication systems for those who are hearing and/or speech impaired. Using an ESP32, flex sensors and an accelerometer, the system interprets the movements of the hands and fingers into either text or voice. This technology allows for enhanced accessibility and participation for individuals who have hearing loss, as well as for the general public. The lightweight, affordable nature of this system makes it an attractive option for daily use, and its advancement in assistive technologies promotes diversity and independence for all.

VI. FUTURE SCOPE

Although the proposed system shows effective performance in recognizing gestures and generating speech output, further improvements can enhance its capability. Future work may focus on training the recognition model with a larger and more diverse dataset to improve accuracy and support additional sign language gestures. Incorporating advanced deep learning models could also strengthen recognition performance under varying lighting and background conditions. In addition, enabling multilingual speech output would allow the system to serve users from different linguistic backgrounds. Integration with mobile devices or cloud platforms may provide greater processing power and remote accessibility. Furthermore, incorporating Internet of Things (IoT) connectivity could allow the device to interact with smart environments, improving its overall usability and scalability.

REFERENCES

[1] Dos Santos, Aline Darc Piculo, Ana Harumi Grota Suzuki, Fausto Orsi Medola, and Atiyeh Vaezipour. "A systematic review of wearable devices for orientation and mobility of adults with visual impairment and blindness." IEEE access 9 (2021): 162306-162324.
[2] Ali, Anum, Priyabrata Parida, Vutha Va, Saifeng Ni, Khuong Nhat Nguyen, Boon Loong Ng, and Jianzhong Charlie Zhang. "End-to-end dynamic gesture recognition using mmWave radar." IEEE Access 10 (2022): 88692-88706.
[3] Kanwal, Tabassum, and Saud Altaf. "Exploring Sensor Fusion Techniques for Enhanced Dynamic Hand Gesture Recognition: A Comprehensive Metadata Analysis." IEEE Sensors Reviews (2025).
[4] Ahmed, Shahzad, and Sung Ho Cho. "Machine learning for healthcare radars: Recent progresses in human vital sign measurement and activity recognition." IEEE Communications Surveys & Tutorials 26, no. 1 (2023): 461-495.
[5] Zabihi, Soheil, Elahe Rahimian, Amir Asif, and Arash Mohammadi. "Trahgr: Transformer for hand gesture recognition via electromyography." IEEE Transactions on Neural Systems and Rehabilitation Engineering 31 (2023): 4211-4224.
[6] Mehta, Amit Sing, Aniket Singh, and Anil Kumar Sagar. "Vision Assist Glasses for Visually Impaired People." In 2024 2nd International Conference on Networking and Communications (ICNWC), pp. 1-8. IEEE, 2024.
[7] Dinelli, Gianmarco, Gabriele Meoni, Emilio Rapuano, Tommaso Pacini, and Luca Fanucci. "MEM-OPT: A scheduling and data re-use system to optimize on-chip memory usage for CNNs on-board FPGAs." IEEE Journal on Emerging and Selected Topics in Circuits and Systems 10, no. 3 (2020): 335-347.
[8] Mehra, Ravish, Owen Brimijoin, Philip Robinson, and Thomas Lunner. "Potential of augmented reality platforms to improve individual hearing aids and to support more ecologically valid research." Ear and hearing 41 (2020): 140S-146S.
[9] Liu, Debang, Tianqi Zhang, Mads Græsbøll Christensen, Chen Yi, and Zeliang An. "Audio- visual fusion with temporal convolutional attention network for speech separation." IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024).
[10] Marakala, Vijaya, G. V. Sriramakrishnan, Geethamanikanta Jakka, Chetan J. Shingadiya, Hesti Prawita Widiastuti, and Geovanny Genaro Reivan Ortiz. "Use of deep learning application in medical devices." In 2022 4th International conference on inventive research in computing applications (ICIRCA), pp. 935-939. IEEE, 2022.
[11] Levine, Sergey. "CAREER: Deep Robotic Learning with Large Datasets: Toward Simple and Reliable Lifelong Learning Frameworks." NSF Award Number 1651843.Directorate for Computer and Information Science and Engineering 16, no. 1651843
[12] Saini, Preeti, Jagpreet Kaur, and Shweta Lamba. "A review on pattern recognition using machine learning." Advances in Mechanical Engineering: Select Proceedings of CAMSE 2020 (2021): 619-627
[13] Krug, Andreas, and Sebastian Stober. "Gradient- Adjusted Neuron Activation Profiles for Comprehensive Introspection of Convolutional Speech Recognition Models." arXiv preprint arXiv:2002.08125 (2020).
[14] Marakala, Vijaya, G. V. Sriramakrishnan, Geethamanikanta Jakka, Chetan J. Shingadiya, Hesti Prawita Widiastuti, and Geovanny Genaro Reivan Ortiz. "Use of deep learning application in medical devices." In 2022 4th International conference on inventive research in computing applications (ICIRCA), pp. 935-939. IEEE, 2022.
[15] Li, Sheng, Jaya Kawale, and Yun Fu. "Deep collaborative filtering via marginalized denoising auto-encoder." In Proceedings of the 24th ACM international on conference on information and knowledge management, pp. 811-820. 2015.
[16] Tan, Ke, Yong Xu, Shi-Xiong Zhang, Meng Yu, and Dong Yu. "Audio-visual speech separation and dereverberation with a two-stage multimodal network." IEEE Journal of Selected Topics in Signal Processing 14, no. 3 (2020):542-553.
[17] Kose, Hatice, and Pinar Uluer. "The uses of technology in L1 and L2/Ln sign language pedagogy." In The Routledge Handbook of Sign Language Pedagogy, pp. 323-338. Routledge, 2019.
[18] Pedersen, Nicolai Fernández. "Audiovisual speech analysis with deep learning." (2021).
[19] Baas, Matthew. "Disentangled Representations in Speech Processing Applications." PhD diss., Stellenbosch University, 2024.
[20] 20. Weng, Juyang. "Symbolic models and emergent models: A review." IEEE Transactions on Autonomous Mental Development 4, no. 1 (2011): 29-53.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY