Arghajeet Dey, 2025, 13:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science,
Engineering and Technology

An Open Access Journal

Gesture Language Translation Using Convolutional Neural Network

Arghajeet Dey, Aniket Chalkapure, Devanshu Naik, Piyush Singh, Salman Mohammedhanif Buddha

Department of Information Technology, Parul Institute of Engineering & Technology,
Faculty of engineering & Technology, Vadodara, Gujarat, India

Abstract- This study presents a revolutionary gesture language translator that can convert sign language into both text and speech output, which can aid in communication for people who are deaf or mute. This method differs from other existing alternatives in that it can convert gestures into text and voice and the other way around, from spoken words into corresponding gestures. By enabling smooth interaction between the hearing and the non-hearing, this dual-mode system improves communication.The system uses a Convolutional neural network (CNN) for gesture recognition and Flask and Python for back-end operations. Google's text- to-speech (TTS) API generates voice output, while Media-pipe is utilized to detect hand landmarks in real-time. In the opposite procedure, vocal inputs are recorded by speech recognition technology and converted into gestures using a pre-stablished sign language model. This product represents a significant achievement in the field of assistive technologies for people with speech and hearing impairments since it has a distinctive feature: two-way translation capabilities.

Keywords: Gesture Language Translator, Sign Language Recognition.

I. INTRODUCTION

Although communication is essential to human connection, it can be extremely difficult for those who are deaf or mute. Conveying ideas and feelings requires the use of sign language, a visual method of communication that makes use of hand gestures, body alignment, and facial expressions. Despite its significance, there is a communication gap between those who are not familiar with sign language and those who make use of it. This gap typically limits social interactions, educational opportunities, and employment involvement for deaf and mute individuals. In order to close this gap, a system that can translate sign language into text and voice and vice versa must be created. This will enable those who can talk to understand sign language as well.

With this effort, a revolutionary bidirectional sign language translator that surpasses existing systems is introduced. Most existing solutions only allow the hearing population to connect with the deaf and mute by translating sign language into text or audio. But in the other direction of communication, this leaves a gap. Our technology provides voice-to-gesture translation, which converts spoken words into corresponding sign language movements, in order to close this gap. Communication in both ways dramatically improves accessibility and creates a welcoming environment where individuals with and without hearing can interact naturally and in real time.

The technological foundation of the system is built on advances in machine learning and computer vision. Convolutional neural networks (CNN's) are used by the system to accurately identify hand gestures for speech and gesture-to-text conversion. Media-pipe powers real-time hand landmark identification, guaranteeing accuracy in distinguishing the different parts of sign language.

After the gestures are identified, textual output is produced, and Google's text-to- speech (TTS) API is used to further convert the written output into speech.

Previously untested is the ability to translate spoken commands into gestures and vice versa. The system records spoken language by utilizing voice recognition technologies, then associates the identified words with pre-stablished sign language motions. The user is subsequently presented with these motions, which enables spoken communication to be depicted through sign language. There are new avenues for contact in social and professional contexts thanks to this special functionality.

The translator's bidirectional nature allows it to be applied to a wide range of use situations. It can help teachers communicate with students who are deaf in educational settings, and it can help hearing and non-hearing colleagues collaborate seamlessly in professional situations. Furthermore, by incorporating this approach into public services, people with speech and hearing impairments can communicate with service providers more successfully.

With the help of real-time gesture and speech processing technology and sophisticated machine learning algorithms, this project seeks to develop a user-friendly, approachable tool that fosters inclusive and closes the communication gap between various groups. In terms of the advancement of assistive technologies for people with hearing and speech impairments, the dual translation capability—gesture-to-text and speech and voice-to-gesture—offers a more extensive and engaging communication platform. In this paper section II discuss background, Section III describes literature survey, section IV methodology and implementation, section V explains the results and discussion and followed by conclusion.

II. BACKGROUND

Computer vision and machine learning techniques have accelerated the advancement of gesture recognition technology. The technological foundation of the system is built on advances in machine learning and computer vision. Convolutional neural networks (CNN's) [18] are used by the system to accurately identify hand gestures for speech and gesture-to-text conversion.

For the deaf and mute communities, sign language is a vital means of communication. However, communication difficulties still exist when engaging with those who do not comprehend sign language. Traditional techniques of interpreting sign language rely on manual interpretation, which can be time-consuming and error-prone. There is an increasing demand for automated solutions that can bridge the communication gap by delivering real- time translation of signs.

III. LITERATURE SURVEY

A. Gesture Recognition Techniques

Convolutional neural networks (CNN's) have been a popular research tool for recognizing American Sign Language (ASL). CNN's [1] to increase real-time performance and get high accuracy. However, its applicability for low-powered devices is limited by its limited vocabulary and high computing needs. Similar to [4] improved recognition accuracy for complicated gestures by using 3D-CNNs to record the spatial and temporal aspects of gestures. Even with these improvements, the method is still computationally demanding and hardware-intensive.[7] improved gesture identification accuracy and robustness by combining deep learning models with image processing approaches for Indian Sign Language recognition. However, there is still a lack of flexibility to various motions and environmental circumstances.

B. Sign Language Translation Systems

The goal of some systems has been to convert sign language into speech or text. A glove-based method for converting Indian Sign Language into speech, "Talking Hands,"[9] Although it showed real-time translating skills, there were issues with its reliance on specialized technology and its limited ability to adjust to complex movements. Similarly, "Mudra," a CNN-based translator [11] specifically for financial services, improved accessibility in particular situations. Its applicability to different signature techniques and other domains was constrained, nevertheless.

C. Voice-to-Gesture Systems

The field of voice-to-gesture translation is yet relatively unexplored. Using video data [16] used Hidden Markov Models (HMMs) to recognize ASL in real time. Although the method effectively handled dynamic motions, it performed poorly in a variety of sign language contexts and was sensitive to video quality. Studies that are now available mostly concentrate on gesture recognition, with little investigation of voice-to-gesture systems. This disparity highlights the uniqueness of our bidirectional translator, which combines gesture synthesis with speech recognition.

There are still large gaps in gesture identification and translation, despite recent advances. Their usefulness is limited by a lack of attention to bidirectional communication systems that include voice-to-gesture and gesture-to- text/voice translation. Real-time deployment on mobile or low-powered devices is hampered by high processing costs and hardware dependencies. Furthermore, limited datasets and vocabularies limit the system's adaptability in real-world scenarios.

By presenting a dual-mode system that can translate from speech to gesture/text and gesture to text/voice, our bidirectional gesture language translator fills these gaps. Real-time performance is guaranteed by using lightweight models and effective algorithms, while system flexibility is increased by having larger vocabularies and the capacity to adapt to different signing styles. This method facilitates smooth real-time communication between hearing and non- hearing people, marking a major advancement in inclusive communication technology.

IV. METHODOLOGY & IMPLEMENTATION

The research methodology involves systematically studying and implementing a Gesture Language Translator using machine learning techniques, focusing on hand gesture recognition. This section describes the research design, data collection methods, system components, model training, and evaluation techniques in detail.

A. Research Design

Creating a real-time hand gesture detection system that can convert hand gestures into text and audio is the main goal of this research. The project employs a quantitative research methodology by evaluating the model's performance using quantifiable data, including evaluation measures, gesture inputs, and model outputs. The implementation of a CNN model [17-18] for hand gesture detection and testing its capacity to convert gestures into relevant text and audio outputs are the main objectives of the experimental design.

The system design combines several essential modules to handle gesture-to-speech conversion and efficiently record, process, and convert gestures into text and audio and vice versa.

Using an external camera or webcam, the input module records hand gestures in real-time, capturing hand movements that are subsequently processed frame by frame. Smooth gesture capture is ensured using high frame rate cameras (30–60 frames per second). To balance real-time processing and gesture recognition accuracy, modifications are made to image resolution and frame size.

According to the schematic of the Gesture Language Translator System, users start the procedure by logging in and registering. Once inside, users are greeted with a main interface that has three primary features: speech recognition, camera activation, and camera closure. The camera recognizes and records motions when it is turned on, and they are then converted into text or voice. As an alternative, users can enable voice recognition by clicking the "Click to Speak" button. This allows spoken instructions to be recognized, processed, and matched to gestures for translation. Voice-to- gesture, gesture-to-voice, and gesture- to-text translation modes may all be switched between with ease using this method.

The processing module, which handles gesture classification and hand landmark detection, is the brains of the system. Using a custom dataset, a pre-trained Convolutional Neural Network (CNN) model extracts important hand landmarks such as joints and fingertips.These landmarks are crucial for differentiating between various movements. For landmark detection, the Media-pipe Hand API is utilized, offering a 21-point skeletal depiction of the hand. For precise gesture detection, this representation incorporates key locations such as the wrist, knuckles, and fingertips. Before being fed into the CNN model for classification, captured images are pre-processed by scaling, normalizing pixel values, and extracting hand landmarks.

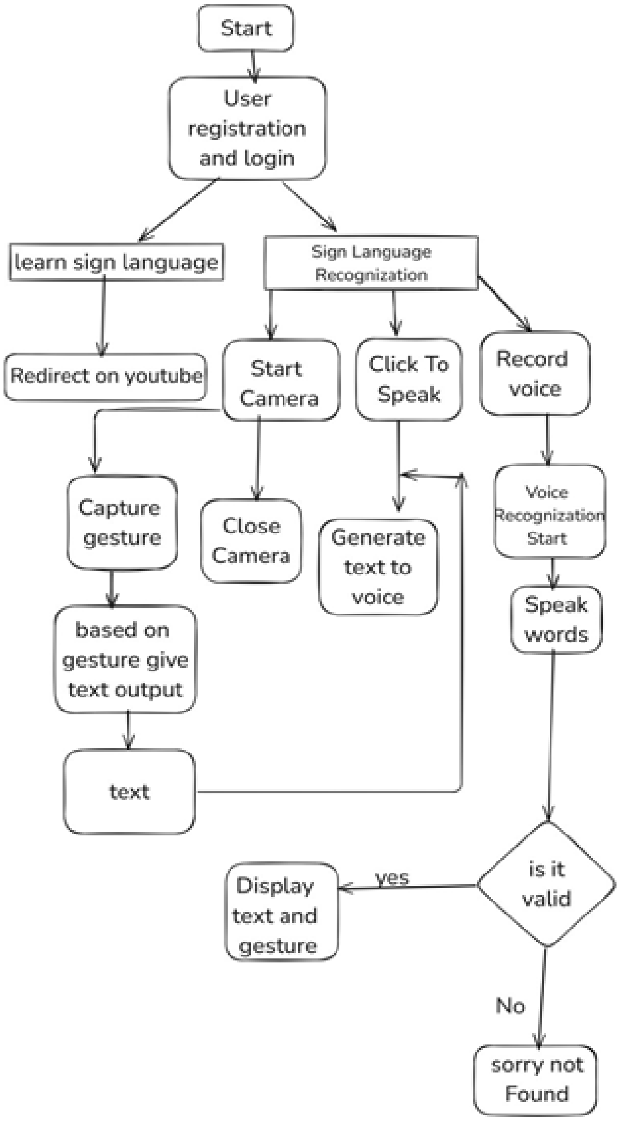


Fig. 1. Workflow of Gesture Language Translation

Gestures are converted into text after classification. After that, a Text-to-Speech (TTS) engine, like Google Text-to- Speech, turns this text into speech and gives the user auditory feedback. With both visual and auditory outputs, the TTS engine reads aloud the text that has been identified in relation to the movements.

In addition to gesture-to-text/speech translation, the system includes a speech-to-gesture conversion feature. This allows spoken words to be recognized and translated into corresponding hand gestures. These gestures can be displayed on-screen, enhancing the interaction by visualizing spoken content. This dual functionality not only supports the translation of gestures into text and speech but also facilitates the translation of speech into gestures, making the system more versatile and interactive.

The architecture ensures that users can interact with the system in a dynamic way, receiving feedback in multiple formats—text, speech, and visual gestures—thereby creating a comprehensive and engaging user experience.

The construction of the gesture recognition system requires the creation of a comprehensive dataset in order to guarantee reliable and accurate model training. The dataset was created by capturing the signs using the webcam/camera. Gestures were captured in a variety of lighting settings to guarantee the dataset's breadth. Participants made motions that represented alphabetic letters and everyday words like "hello," "what," and "how."

The CNN model used to classify gestures was trained on this pre-processed dataset. The model design employed many Convolutional layers to extract spatial features from hand landmarks, followed by the addition of max pooling layers to down sample feature maps. Next, classifiers were fully connected layers. Each class of gestures had probabilities provided by a soft-max output layer, and the gesture with the highest probability was selected as the expected gesture.

A variety of optimization techniques were needed for the model's training. Because of its high efficiency, the Adam optimizer was utilized, and categorical cross-entropy was the loss function used. To hyper parameter tune, grid search and cross-validation were employed to find the optimal combination of kernel size, learning rate, and number of Convolutional filters.

The model's performance was evaluated using metrics such as confusion matrix. The confusion matrix provided insight into the performance for each gesture class by highlighting classifications and areas in need of improvement.

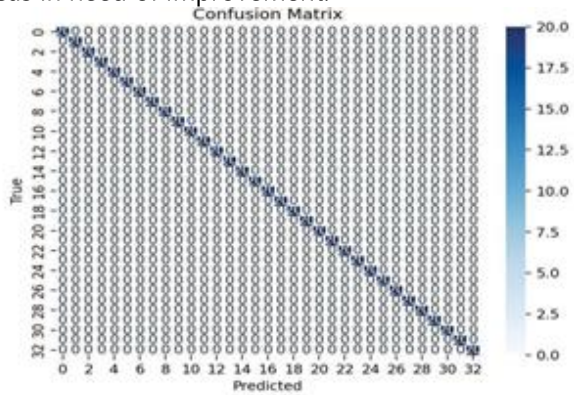


Fig. 2. Confusion Matrix of Predicted outcome versus Actual outcome

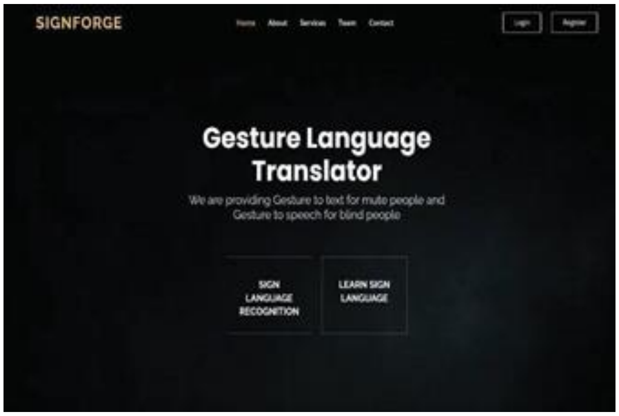
The Text-to-Speech (TTS) engine was used in the audio feedback component. The TTS engine translated motions into text and then audibly said the text, so that users could hear the translated sentence in addition to seeing it on screen.

The system's core functionality includes gesture-to-text and gesture-to-speech translation, supported by a Text-to- Speech (TTS) engine that provides auditory feedback. Additionally, the system features voice-to-gesture translation. This innovative functionality allows spoken words to be recognized and translated into corresponding hand gestures, which can be displayed on-screen. This dual capability— converting both gestures to text/speech and speech to gestures—enhances the system's versatility and interactive experience.

Thorough testing was done to verify the system. The system's ability to accurately recognize gestures on the fly as well as the response time from gesture input to text or voice output were used to measure the system's real-time performance. Diverse sign language proficiency levels participated in user testing to provide input on usability, accuracy, and overall experience. Error analysis was used to identify typical problems, such as unclear hand motions or postures, and temporal filtering and smoothing were applied to improve accuracy and robustness.

V. RESULTS AND DISCUSSIONS

To help people who use sign language and those who are not familiar with it to communicate using sign language , the Gesture Language Translator system is created. This software enables smooth interaction and encourages inclusiveness by converting hand motions into written and auditory outputs. A wide range of hand motions that were mapped to frequently used words and phrases in sign language were utilized to evaluate the system.
This chapter explores the Signforge website, the main platform for the Gesture Language Translator, including its features, implementation process, and practical use.



Gesture Language Translator

We are providing Gesture to text for mute people and Gesture to speech for blind people

Fig. 3. Home Page of Gesture Language Translator

HomePage Overview:
The Signforge homepage serves as a starting point to explore the functionalities of website, providing clean and user-friendly interface. Users are served with two core features:

1. **Sign Language Recognition:** Aiming at real-time interaction, this feature allows users to recognize sign language and translating them into corresponding text and audio outputs.
2. **Learn Sign Language:** This feature offers tutorial for those who don't know sign language and users can learn by watching those tutorials.

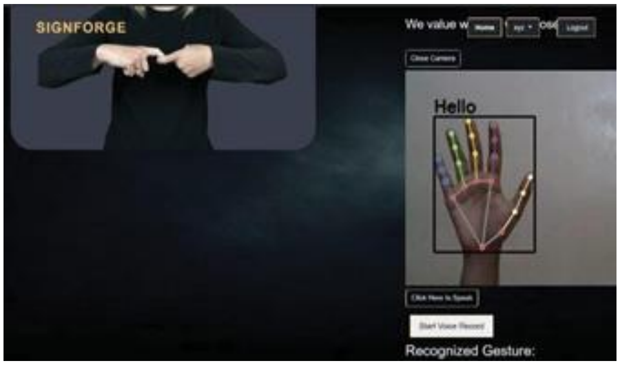


Fig. 4. Sign to text or speech conversion

Sign Recognition Page:
By clicking on Sign Language Recognition option, users are redirected to a dedicated page that allows users to capture their gestures and can translate it to corresponding text and audio

This page offers two functionalities:

1. **Start Camera:**
It activates the user's webcam to capture hand gestures. The gestures captured are processed by Convolutional Neural Network (CNN) model trained to recognize gestures. The system translates the gestures into text, which is displayed on screen.
2. **Click Here to Speak:**
It converts the gestures text output into speech. It enables users to hear the audio of translated words, making it accessible to those who prefer auditory feedback or have visual impairments.

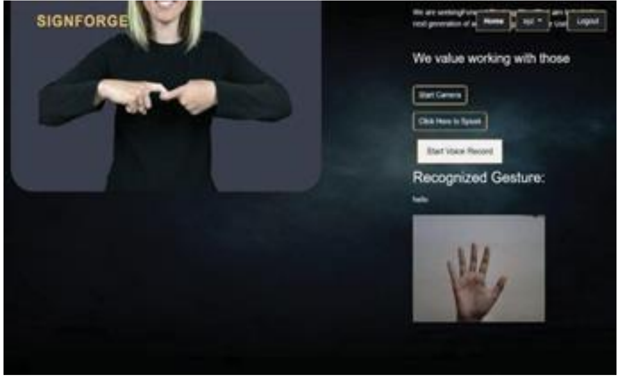


Fig. 5. Voice To Gesture conversion

Voice-to-Gesture Translation:
In addition to translating gestures into text and audio, the system provides the reverse functionality too, converting spoken words into corresponding gestures.

Speech Recognition Integration:
The voice-to-gesture feature utilizes Google Speech Recognition API to process spoken words

1. **Input Processing:**
The system captures the audio input directly through microphone and converts it into text using speech recognition API
2. **Gesture Mapping:**
The processed text is matched against a predefined dataset of gestures. The corresponding gesture is displayed on screen as a static image.

VI. CONCLUSION

An important development in the fusion of real-time communication technology and machine learning is the gesture language translator system. Through the integration of voice-to-gesture conversion and gesture-to-text/speech translation, the system provides an adaptable platform for both practical and instructional uses. Users may interact with the technology in a smooth and practical way thanks to its novel voice-to-gesture capability and real-time hand gesture detection.

The system's efficiency is highlighted by its comprehensive architecture, which includes a large dataset, complex model training, and real-time processing capabilities. Instructive visuals and other user-friendly aspects improve usability even further by helping users maximize gesture recognition. With the use of these functions, the system helps individuals who use sign language communicate more effectively and supports educational initiatives by making sign language learning easier.

Furthermore, the actual implementation of the system will depend on optimizing its real-time processing speed and accuracy through hardware advancements and optimization approaches. Creating a mobile application could improve accessibility by enabling users to utilize tablets or smartphones to engage with the system.

In-depth usability research and the exploration of user feedback will yield insightful information that can be used to improve the interface and user experience as a whole. Lastly, looking into interaction with other platforms and assistive technologies could broaden the system's impact and reach, making it a more adaptable tool for instruction and communication in a variety of settings.

By tackling these issues, the system can advance further, providing more assistance for sign language interaction and fostering an inclusive digital environment.

REFERENCES

1 Bantupalli, K., Xie, Y.: American Sign Language Recognition using Deep Learning and Computer Vision. Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018. 4896–4899 (2019).

2 Zhang, J., Li, M., and Li, Y. (2019). a study on hand gesture identification. 20(6), 763-776, Frontiers of Information Technology & Electronic Engineering.

3 Jaiprakash Narain Dwivedi, "Hierarchical Classification Based on Conditional SOM for Situation Analysis", International Journal of Grid and Distributed Computing, ISSN: 2005-4262, Vol. 13, No. 1, (2020), pp. 594-605.

4 Al-Hammadi, M., Muhammad, G., Abdul, W., Alsulaiman, M., Bencherif, M. A., & Mekhtiche, M. A. (2020). Hand gesture recognition for sign language using 3DCNN. IEEE Access, 8, 79491-79509.

5 Sharma, A., Sharma, N., Saxena, Y., Singh, A., Sadhya, D.: Bench marking deep neural network approaches for Indian Sign Language recognition. Neural Compute. Appl. 2020 3312. 33, 6685–6696 (2020).

6 Jaiprakash Narain Dwivedi, "Conditional Self-Organizing Map for Hierarchical Classification in Road-Vehicle Situation", International Journal of Advanced Science and Technology, ISSN: 2005-4238, Vol. 29, No. 3,(2020),pp.31313141.

7 Neel Kamal Bhagat, Vishnusai Y, Rathna G N, "Indian Sign Language Gesture Recognition using Image Processing and Deep Learning," 2019 Digital Image Computing: Techniques and Applications (DICTA). doi:10.1109/dicta47822.2019.8945850.

8 Jaiprakash Narain Dwivedi, "Analyzing the Collision Risk of Driver- less Vehicle Based on Clustering of shapes of the road", International Journal of Control and Automation, ISSN 2005-4297, Vol. 13, No.1, (2020), pp. 01-10.

9 [9] S Yarisha Heera, Madhuri K Murthy, Sravanti V S, "Talking hands — An Indian sign language to speech translating gloves," 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), IEEE, 2017, pp. 746-751.

10 [10] Jaiprakash Narain Dwivedi, "Situation, Scene and Scenario Classifications and Understanding" in International Journal of Advance Research in Science and Engineering, ISSN: 2319-8354, Volume 06, Special Issue No. 01, Dec 2017, 972-987.

11 Jayadeep, G.; Vishnupriya, N.V.; Venugopal, V.; Vishnu, S.; Geetha, M. Mudra: Convolutional neural network based Indian sign language translator for banks. In Proceedings of the 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 13–15 May 2020; pp. 1228–1232.

12 Jaiprakash Narain Dwivedi, "The possibility of existence of life after death", in International Journal of Science Technology and Management, ISSN(o) 2394-1537, ISSN (P) 2394-1529, Volume 06, Issue 05, May 2017, 417-421.

13 Clebeson Canuto dos Santos, Jorge Leonid Aching Samatelo, Raquel Frizera Vassallo, "Dynamicgesture recognition by using CNNs and star RGB: A temporal information condensation,"Neurocomputing, Volume 400, 2020, Pages 238-254, ISSN 0925-2312.

14 Jaiprakash Narain Dwivedi, "Estimation of the Collision Risk of Autonomous Vehicle using Clustering", 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), 22-23 March 2019
, IEEE(ISBN:978-1-5386-81909),DOI:10.1109/i-PACT44901.2019.8960025

15 Mahidhar, B. V. S., Sankeerthana, D. L., Reddy, K. B., Nikhath, G. A., & Poovaraghan, R. J.MEDICAL TRANSCRIPTION USING SPEECH RECOGNIZER.

16 Starner, T., Weaver, J., and Pentland, A. (1998. Using hidden Markov models, real-time recognition of American Sign Language from video. IEEE Transactions on, Pattern Analysis and Machine Intelligence, 20(12), 1371–1375.

17 S. Venkatasubramanian, Jaiprakash Narain Dwivedi, S. Raja, N. Rajeswari, J. Logeshwaran, Avvaru Praveen Kumar, "Prediction of Alzheimer’s Disease Using DHO-Based Pre-trained CNN Model", Mathematical Problems in Engineering, ISSN: 1563-5147, vol. 2023, Article ID 1110500, 11 pages, 2023.

18 Zhao, X., Wang, L., Zhang, Y. et al. A review of Convolutional neural networks in computer vision.Artif Intell Rev 57, 99 (2024).