2024 IEEE International Conference on Future Machine Learning and Data Science (FMLDS) | 979-8-3503-9121-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/FMLDS63805.2024.00097

Enhancing Sign Language Interpretation with Multi-headed CNN,Hand Landmarks and Large Language Model(LLM)

Chaitanya Kakade
University of Mumbai
Mumbai, India
kakadechaitanya77@gmail.com

Nidhi Kadam
University of Mumbai
Mumbai, India
nidhikadam68@gmail.com

Vishal Kaira
University of Mumbai
Mumbai, India
kairavishal37@gmail.com

Rishi Kewalya
University of Mumbai
Mumbai, India
rishikewalya@gmail.com

***Abstract*—Sign language is an important mode of communication for the deaf and mute community. Despite its importance, there is still a large communication gap between deaf community and the hearing world. We introduce a new system that converts sign language into text, using a novel multiheaded Convolutional Neural Network (CNN) that is trained on three different sets of images-raw images, uniquely segmented images, and hand landmarks information simultaneously to recognize sign language gestures accurately for different hand textures under disparate background conditions. Additionally, a Large Language Model (LLM) is incorporated to transform these recognized signs into concise and meaningful sentences. This combination ensures the text is both understandable and grammatically correct, thereby reducing the communication delay that exists in multiple Sign Language Recognition (SLR) systems. The model excels in capturing nuanced features and variations in sign gestures. Tests show that our novel approach is highly effective in real-time recognition, demonstrating its potential to improve communication for the Deaf community greatly.**

Keywords—American Sign Language (ASL), Convolutional Neural Network (CNN), Large Language Model (LLM), hand landmarks, Natural Language Processing (NLP).

I. INTRODUCTION

Deafness and hearing loss is widespread and found in every region and country. According to [8]., more than 1.5 billion people suffer from hearing loss (nearly 20% of the global population), 430 million of them have disabling hearing loss and it is expected that by the year 2050, there is a high chance of having more than 700 million people with disabling hearing loss. With an increasing proportion of deaf people, there is an urgent need to develop systems that can bridge the communication gap between these individuals and non-deaf population. Sign language is a predominant method of communication for millions of deaf individuals worldwide. It is a combination of multiple hand gestures, facial expressions, and bodily movements to convey rich and complex linguistic information. Despite its importance, sign language users often face significant communication barriers when interacting with individuals who do not understand sign language. This communication gap can lead to social isolation, reduced access to services, and limited opportunities for education and employment for Deaf individuals as per [12]. Napier et al. (2006).

In the recent years, adoption of deep learning approaches have increased due to its ability in achieving outstanding results on several complex cognitive tasks and its ability to learn and understand massive amounts of data as per [3]. Alzubaidi et al. (2021). CNNs have demonstrated remarkable success in image and video recognition tasks, making them well-suited for recognizing the intricate details of hand gestures and movements involved in sign language [18]. Simonyan et al. (2014). However, accurately translating these gestures into coherent and meaningful sentences remains a significant challenge. Also, the performance of CNN improves in accurately detecting features of certain parts of an image when the background is removed. [11]. Liang et al.(2024).To address this issue, we have developed a novel multiheaded CNN architecture capable of recognizing sign language gestures with high accuracy. Furthermore, we have integrated a Large Language Model (LLM) to convert the recognized gestures into coherent sentences. This integration ensures that the translated text is not only accurate but also contextually and grammatically correct. Previous studies have highlighted the potential of combining CNNs with language models for sign language recognition, but our approach represents a significant advancement in both accuracy and usability.

We have also utilised hand landmarks information to train the proposed model. It provides important information such as spatial coordinates of hands [20]. Tomasz et al. (2016), which helps in devising distances and angles between various points on hands, that serves as a reliable information for extracting features. We also provide detailed experimental results demonstrating the effectiveness of our approach compared to existing methods. Our findings indicate that this system offers a practical solution for improving communication, thereby contributing to a more inclusive society.

979-8-3503-9121-3/24/$31.00 ©2024 IEEE
DOI 10.1109/FMLDS63805.2024.00097

II. LITERATURE REVIEW

With the advent of deep learning, CNNs have become a popular choice for sign language recognition due to their superior performance in image and video processing tasks. For instance, [15]. Pigou et al. (2015) proposed a CNN-based approach for isolated sign language recognition, achieving notable improvements in accuracy. Their model utilized spatial-temporal features to capture the dynamic nature of sign language, addressing some of the limitations of earlier methods. Despite these advancements, traditional CNN architectures often struggle with variations in background and lighting conditions, which can significantly impact recognition accuracy. To mitigate these challenges, [9]. Dong et al. (2015) introduced a robust CNN model that incorporated hand shape and motion information. Their approach demonstrated enhanced resilience to environmental variations, making it more suitable for real-world applications.

Hand landmark detection has also emerged as a critical component in improving the accuracy of sign language recognition systems as seen in [14]. Priya et al. and [6, 21]. Zhang et al. (2017), developed a framework that combined CNNs with a pose estimation algorithm to accurately detect hand landmarks. This method enabled more precise gesture recognition by focusing on the key features of hand movements, significantly reducing the error rates in sign language translation. Moreover, LLMs have shown immense potential in natural language processing tasks. LLMs, such as GPT-3 [4]. Brown et al. (2020), have revolutionized the field by demonstrating the ability to generate human-like text and understand complex linguistic contexts. These models utilize large amounts of data and sophisticated architectures to achieve high levels of fluency and coherence in text generation. Furthermore, Microsoft and Leap Motion have developed distinct methods to detect and monitor hand and body motions through their respective devices, Kinect [10]. Huang et al. (2011) and the Leap Motion Controller (LMC). Kinect identifies body skeletons and tracks hand movements, while LMC uses built-in cameras and infrared sensors to differentiate and track hands. Utilizing this technology, [19]. Sykora et al. (2014) employed the Kinect system to capture depth data from 10 hand gestures. They used the speeded-up robust features (SURF) technique for classification, achieving an accuracy of 82.8%. However, this approach has limitations, as it has not been tested on a larger database and may be non-invariant to gesture orientation due to the use of modified feature extraction methods like SIFT and SURF [17]. Shanta et al. (2018) Also, these devices are quite expensive and non-portable which adds another layer of complexity in the process of fluent and swift communications. Many works such as [16]. Refat et al. (2023) have been done, which uses multi-headed CNN architecture.

III. METHODOLOGY

The overall process for developing a reliable sign language to voice converter can be depicted in figure 5. We propose a novel architecture of a multi-headed CNN [13]. Pathan et al. (2022) that aims to increase the accuracy and solve background constraints faced by typical CNN. Every input frame captured goes through a series of novel image segmentation process, extracting essential features in branch 1 of the proposed architecture. The system also extracts Hand Landmark points from the frame and passes the information to the model, thereby classifying the image to belong to one of the 26 categories of the English alphabet A-Z. Moreover, the classified characters are stored until they form a word that gets autocorrected using methods described in [2]. Aliguliyev et al. (2009). After the accumulation of certain words, the system transfers the tokens to a local LLM provided by [7]. Chung et al. (2022) that converts these words into coherent and meaningful sentences as per [5]. Bumgardner et al.(2024), thereby eliminating the need for a deaf individual to use stop words or complete sentences in order to interact with the system and hence improving the speed and overall ease of communication.

A. Dataset

Due to the challenges faced by utilizing the available data repository [1]., such as invalid data, inappropriate background conditions and poor lighting conditions that hampered the ability of any computer vision model to extract essential and non-negotiable features, we gathered a high quality validated American Sign Language dataset. ~1015 images were captured for each of the English alphabets A-Z, taken under different variations and lighting conditions as depicted in figure 1. Particularly, a high contrast was maintained between the hands and the background for efficient feature extraction by the proposed neural network architecture. Furthermore, we tried to vary the resolution in order to add another layer of complexity in our dataset.

Fig. 1. American Sign Language Dataset

B. Data Preprocessing

A novel image processing technique has been proposed that is able to reliably distinguish hand edges from the background that surpasses the direct conventional methods such as canny edge detection in terms of direct detection of hand edges, ensuring superior accuracy in identifying distinguishing features, particularly in challenging scenarios such as varying background conditions. This advanced approach along with the conversion of raw images into grayscale, followed by the application of a Gaussian filter to remove noise, helped in subsequent processing.

A significant enhancement was made by making use of adaptive thresholding, a technique that dynamically determines optimal thresholds for binarization based on local image characteristics. This contributed to the accuracy of feature extraction. Next, we applied Otsu's thresholding technique to further refine the segmentation process. Otsu's method automatically calculates the optimal threshold to maximize inter-class variance, effectively separating the hand from the background with remarkable precision.

An important step in the process involved inverted thresholding, which inverted the binary image obtained from the thresholding operations. The inversion facilitated the isolation of the hand and nails, rendering them as predominant features, thus enhancing their visibility for subsequent analysis.

The integration of the proposed image processing pipeline resulted in a highly accurate and robust method for identifying nails and tracing the hand with exceptional precision, even in the presence of diverse background conditions. As a result, the integration of grayscale conversion, Gaussian filtering, adaptive thresholding, Otsu's thresholding, and inverted thresholding represents a novel image processing pipeline as seen in figure 2. Moreover, we extracted hand landmark coordinates to be used in branch 3 of the proposed model, later discussed in section D. Applied process can be seen in figure 3 and hand-landmark points in figure 4.

Data augmentation was incorporated to increase the size of the dataset and enhance the model training. Images were rotated up to 10 degrees, shifted horizontally and vertically by up to 20%. Image flipping was not performed, either horizontally or vertically as it changes the semantics of the gesture.

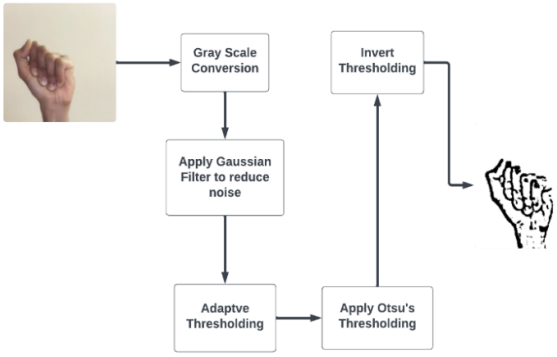

Fig. 2. Image Segmentation process flowchart

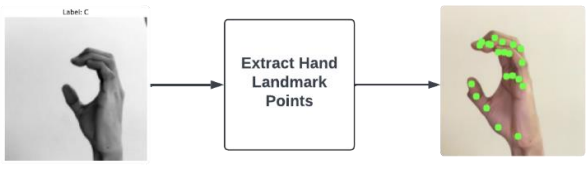

Fig3. Hand-Landmark point extraction

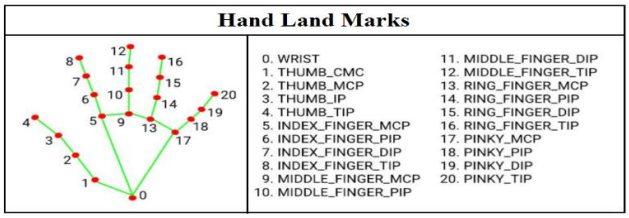

Fig. 4. Hand landmark points

C. Model Architecture

The Proposed multi-headed CNN Architecture is designed to tackle the challenges of recognizing hand gestures in diverse conditions, such as different hand sizes, colours, presence of ornaments, backgrounds, and lighting conditions. This architecture uses a unique 3-branch parallel CNN approach to achieve high accuracy. The entire process is divided into three main parts, one is the raw image processing, another one is the image segmentation processing and last one is the hand landmarks extraction. After the individual processing had been completed, a multi-headed CNN model was built to train on these data. Before processing through a fully connected layer for classification, we merged all three channel's features so that the model could choose between the best weights. This working procedure is illustrated in Fig. 6.

D. Branch 1

The branch 1 of the model is trained on raw images of size 128 by 128 pixels in dimensions with a single grey-scale channel and follows a sequential approach, building layer by layer. The network starts by initializing a sequential model. The first layer is a 2D convolutional layer with 64 filters of size 3x3, applied to an input image. Padding is set to 'same' to ensure the output dimensions match the input dimensions. Batch normalization is then applied to normalize the inputs of the activation function, which helps in stabilizing and accelerating the training process. A ReLU (Rectified Linear Unit) activation function is used to introduce non-linearity, followed by a Max-Pooling layer which helps in reducing the spatial dimensions by a factor of two, and a Dropout layer set to a suitable rate of 25% to prevent any hidden and unnoticed overfitting.

This pattern is repeated with increasing complexity in the subsequent layers. After the series of convolutional layers, the model includes a Flatten layer to convert the 2D matrix into a 1D vector. This is followed by two fully connected (dense) layers with 256 and 512 neurons respectively, each followed by batch normalization, ReLU activation, and dropout. The final layer is a dense layer with 26 neurons (corresponding to the 26 classes) and a SoftMax activation function, which outputs a probability distribution over the classes. The model is compiled using the Adam optimizer with a learning rate of 0.0001, and the loss function used is sparse categorical cross-entropy.

E. Branch 2

The input to the second branch is the segmented image of hand sign as per the segmentation process depicted in figure 2. The network consists of several convolutional layers, each followed by batch normalization, ReLU activation, max-pooling, and dropout layers. The first set of convolutional layers

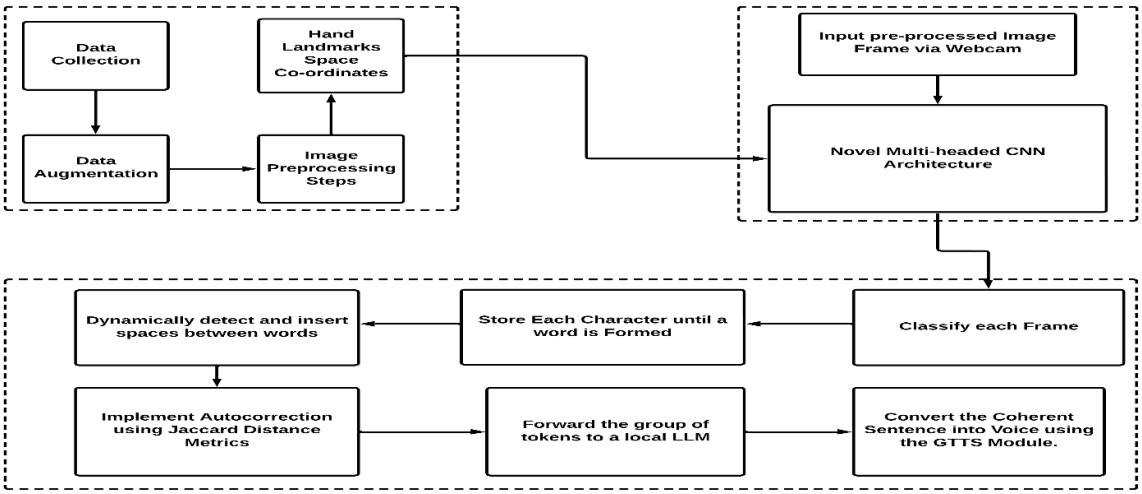

Fig 5. Overall Architecture of the proposed system

contains 64 filters of size 3x3, followed by batch normalization, another 64-filter convolutional layer, batch normalization, max-pooling, and dropout with a rate of 25%. The second set of convolutional layers contains 128 filters, following the same pattern as the first set. The third set includes 256 filters, again following the same pattern. These convolutional layers are responsible for extracting hierarchical features from the input images. Following the convolutional layers, the model includes a flatten layer to convert the 3D feature maps into 1D vectors, which are then fed into fully connected (dense) layers. The first dense layer has 512 neurons with ReLU activation, followed by batch normalization and a dropout rate of 50%. The second dense layer has 256 neurons, also with ReLU activation, batch normalization, and a 50% dropout rate. The final layer is a dense layer with a number of neurons equal to 26, using the SoftMax activation function to output a probability distribution over the classes. The model is compiled using the Adam optimizer and categorical cross-entropy loss, which is appropriate for multi-class classification.

F. Branch 3

The third branch uses hand landmarks information as an input, consisting of 21 key points of the hand, such as fingertips, knuckles, and points on the palm. Unlike the other two branches that use image data, this branch makes use of geometric information from these landmarks. The Z index, which indicates the depth or distance of each landmark from the camera, is kept constant. This consistency helps the model accurately recognize hand gestures despite variations in hand size and orientation. The network begins with an input layer that accepts data with a dimension of 21x2 representing the coordinates of 21 hand landmarks, each having an x and y value. The first hidden layer is a dense layer consisting of 128 neurons and utilizing the ReLU activation function. This is followed by a dropout layer with a rate of 20%. The model includes a second dense layer with 64 neurons, using a ReLU activation function followed by another dropout layer with a higher rate of 30%, providing additional regularization. The final layer uses the SoftMax activation function to convert the output into a probability distribution over the classes. Two callbacks are employed to enhance the training process model checkpoint and early stopping.

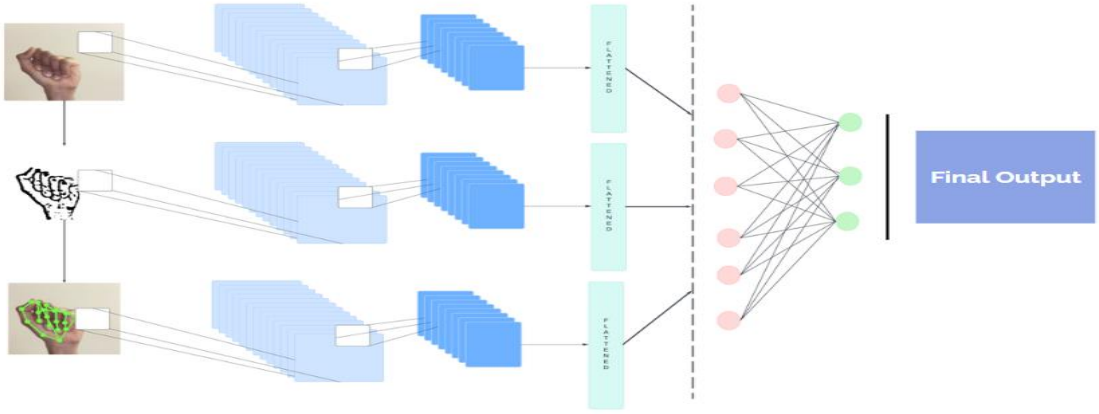

Fig 6. Multi-Headed CNN Architecture with three parallel branches

IV. RESULT AND ANALYSIS

We achieved an accuracy of 99.16%, 99.49% and 98.54% while training and 99.01%, 98.17%, and 98.61% while testing for branch 1, branch 2, and branch 3 of the proposed multi-headed CNN architecture respectively as shown in detail in table 1. During classification task, if one layer is gave a result with less accuracy, it could be complemented by the other layer's weight, and it is possible that combining both results could provide a positive outcome. We used this theory and successfully improved the final validation and test results. The combined approach, which incorporates raw image, segmented image and hand landmark information, is effective for the task when accuracy is priority. On the other hand, branch 3, despite having fewer parameters and lower memory consumption, performed impressively well. Accuracy and loss curves for all the branches can be seen in figures 7 to 11.

TABLE I. ANALYSIS OF ALL THE BRANCHES OF MULTI-HEADED CNN

CNN Branch	*Validation Accuracy*	*Validation Loss*	*Train Accuracy*	*Test Accuracy*
Branch 1	99.34%	0.097%	99.16%	99.01%
Branch 2	98.18%	4.63%	99.49%	98.17%
Branch 3	99.95%	2.67%	98.54%	98.61%

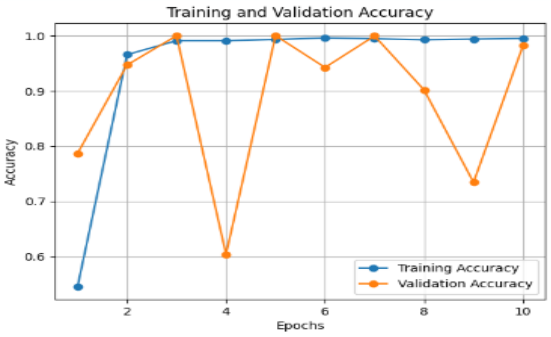

Fig 7. Training and validation accuracy graph for branch 1 of proposed multi-headed CNN

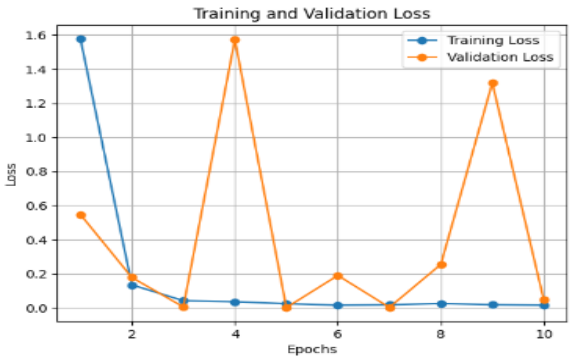

Fig 8. Training and validation loss graph for branch 1 of proposed multi-headed CNN

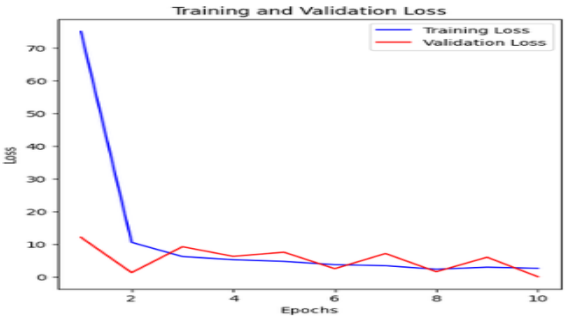

Fig 9. Training and validation accuracy graph for branch 2 of proposed multi-headed CNN

Fig 10. Training and validation loss graph for branch 2 of proposed multi-headed CNN

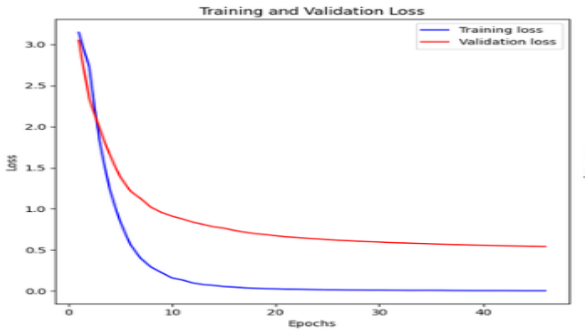

Fig 11. Training and validation accuracy graph for branch 3 of proposed multi-headed CNN

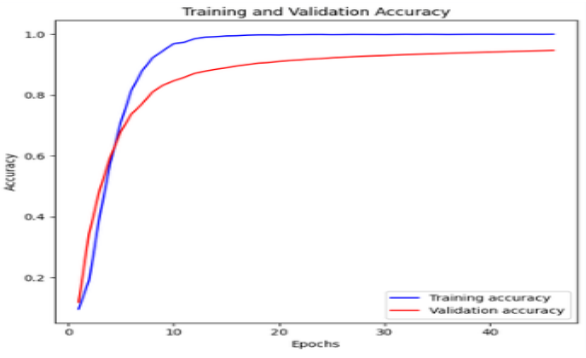

Fig 12. Training and validation loss graph for branch 3 of proposed multi-headed CNN

V. CONCLUSION

By developing a novel multi-headed CNN architecture, we were able to achieve high accuracy, overcome major background constraints and reduce the delay in time in terms of gesture detection. Additionally, a novel technique was introduced which integrated the system with an LLM. This innovative approach addressed the nuances of sign language translation by ensuring that the generated sentences not only maintained accuracy but also adhered to grammatical rules which enhanced the coherence and meaningfulness of the translated sign language gestures in real-time.

REFERENCES

[1] Akash Nagaraj. (2018). ASL Alphabet [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/29550

[2] Aliguliyev, R. M. (2009). A new sentence similarity measure and sentence based extractive technique for automatic text summarization. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2008.11.022.

[3] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8

[4] Brown, Tom & Mann, Benjamin & Ryder, Nick & Subbiah, Melanie & Kaplan, Jared & Dhariwal, Prafulla & Neelakantan, Arvind & Shyam, Pranav & Sastry, Girish & Askell, Amanda & Agarwal, Sandhini & Herbert-Voss, Ariel & Krueger, Gretchen & Henighan, Tom & Child, Rewon & Ramesh, Aditya & Ziegler, Daniel & Wu, Jeffrey & Winter, Clemens & Amodei, Dario. (2020). Language Models are Few-Shot Learners.

[5] Bumgardner VKC, Mullen A, Armstrong SE, Hickey C, Marek V, Talbert J. Local Large Language Models for Complex Structured Tasks. AMIA Jt Summits Transl Sci Proc. 2024 May 31;2024:105-114. PMID: 38827047; PMCID: PMC11141822.

[6] Chen, Yiqing. (2019). Comment on the work of Zhang et al. (2017, Journal of Inequalities and Applications). Journal of Inequalities and Applications. 2019. 10.1186/s13660-019-2142-3.

[7] Chung, H., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q., & Wei, J.. (2022). Scaling Instruction-Finetuned Language Models.

[8] Deafness and hearing loss (2024) World Health Organization. Retrieved from https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss (Accessed: 02 March 2024).

[9] Dong, C., Leu, M. C., & Yin, Z. (2015). American sign language alphabet recognition using microsoft kinect. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 44-52).

[10] Huang, F., & Huang, S. Interpreting american sign language with Kinect. Journal of Deaf Studies and Deaf Education, [Oxford University Press], (2011).

[11] Liang, J., Liu, Y. and Vlassov, V. (2024) The impact of background removal on performance of Neural Networks for Fashion Image Classification and Segmentation.

[12] Napier, J., McKee, R., & Goswell, D. (2006). Sign Language Interpreting: Theory and Practice in Australia and New Zealand. Federation Press.

[13] Pathan, R. K. et al. Breast cancer classification by using multi-headed convolutional neural network modeling. Healthcare 10(12), 2367. https:// doi. org/ 10. 3390/ healthcare 10122 367 (2022).

[14] Priya, & B. J. Sandesh (2023). Hand Landmark Distance Based Sign Language Recognition using MediaPipe.

[15] Pigou, L., Dieleman, S., Kindermans, P. J., & Schrauwen, B. (2015). Sign language recognition using convolutional neural networks. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13* (pp. 572-578). Springer International Publishing.

[16] Refat Khan Pathan, Munmun Biswas, Suraiya Yasmin, Mayeen Uddin Khandaker, Mohammad Salman, & Ahmed A.F. Youssef (2023). Sign language recognition using the fusion of image and hand landmarks through multi-headed convolutional neural network. Scientific Reports, 13.

[17] Shanta, S. S., Anwar, S. T., & Kabir, M. R. Bangla Sign Language Detection Using SIFT and CNN. In 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 1–6 (IEEE, 2018). https:// doi. org/ 10. 1109/ ICCCNT.2018. 84939 15.

[18] Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.

[19] Sykora, P., Kamencay, P. & Hudec, R. Comparison of SIFT and SURF methods for use on hand gesture recognition based on depth map. AASRI Proc. 9, 19–24. https:// doi. org/ 10. 1016/j. aasri. 2014. 09. 005 (2014).

[20] Tomasz Grzejszczak, Michal Kawulok, & Adam Galuszka (2016). Hand landmarks detection and localization in color images. Multimedia Tools and Applications, 75.

[21] Zhang, K., Zuo, W., Gu, S., & Zhang, L. (2017). Learning deep CNN denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3929-3938).