2024 4th Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 23 - 25, 2024

Gesture to Text Recognition using Deep Learning Approach with VGG16 and ResNet50 for Sign Language

Jatin Sharma[1]
Chitkara University Institute of Engineering and Technology, Chitkara University,
Punjab, India
jatin.1003@chitkara.edu.in

Kanwarpartap Singh Gill[2]
Chitkara University Institute of Engineering and Technology, Chitkara University,
Punjab, India
kanwarpartap.gill@chitkara.edu.in

Mukesh Kumar[3]
Computer Science & Engineering, Graphic Era Hill University,
Dehradun, Uttarakhand, India, 248002
kumarsidcse@gmail.com

Ruchira Rawat[4]
Computer Science & Engineering, Graphic Era Deemed to be University,
Dehradun, Uttarakhand, India, 248002
ruchira.rawat.cse@geu.ac.in

2024 4th Asian Conference on Innovation in Technology (ASIANCON) | 979-8-3503-5421-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ASIANCON62057.2024.10838041

***Abstract*— Sign language recognition (SLR) plays a crucial role in enabling communication for individuals who are deaf or hard of hearing. This study explores the application of two prominent deep learning models, VGG16 and ResNet50, in the context of SLR tasks. By employing these sophisticated architectures, we achieved remarkable accuracy rates of 99.92% with VGG16 and 99.95% with ResNet50 in identifying sign language gestures. Our research highlights the exceptional performance of these models in interpreting hand movements and gestures with high precision, thereby enhancing communication for sign language users. Through the application of cutting-edge deep learning methods, this study contributes to the ongoing advancement of SLR systems, offering exciting prospects for fostering inclusive communication and accessibility.**

Keywords*— *Sign language recognition, Deep learning, VGG16, ResNet50, Gesture recognition.

I. INTRODUCTION

The introduction frames our exploration into American Sign Language (ASL) recognition, highlighting its crucial role in fostering inclusive communication within the deaf and hard-of-hearing community. ASL, which employs hand gestures, facial expressions, and body movements to convey messages, is the primary language for an estimated 70 million deaf individuals globally. Despite its significance, traditional communication methods present challenges, underscoring the need for technological advancements in ASL recognition. Recent breakthroughs in deep learning have revolutionized this field, offering powerful tools for highly accurate interpretation of sign language gestures. Our objective is to develop and evaluate ASL recognition models using two prominent architectures: ResNet50 and VGG16. By leveraging deep learning, we aim to create reliable and accessible ASL recognition systems, facilitating seamless and inclusive communication for sign language users.

A. VGG16 Overview

VGG16 is a convolutional neural network (CNN) architecture that achieved top honors in the ILSVRC (ImageNet) competition in 2014. It remains a distinguished vision model architecture due to its design simplicity and effectiveness. The key feature of VGG16 is its consistent use of 3x3 convolutional filters with a stride of 1, accompanied by same padding and 2x2 max pooling layers with a stride of 2. This uniform arrangement of convolutional and pooling layers extends throughout the network. At its conclusion, VGG16 incorporates two fully connected (FC) layers followed by a softmax layer for output classification. The "16" in VGG16 denotes the presence of 16 layers with learnable weights. This extensive network comprises approximately 138 million parameters, making it a substantial and influential model in the realm of deep learning.

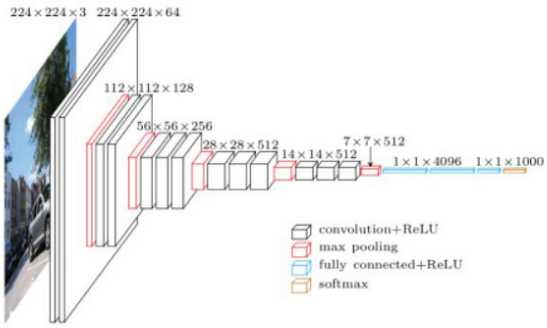

Fig. 1. Architecture of VGG16

B. ResNet50

The ResNet50 model represents a major leap in convolutional neural network (CNN) development, particularly for training deeper networks. Introduced by Microsoft Research in 2015, ResNet50 brought a groundbreaking approach with its implementation of residual connections. These connections allow for the training of significantly deeper neural networks with improved accuracy and optimization. One of the major challenges in deep network training is the vanishing gradient problem, where gradients diminish and hinder the learning process as they propagate through the layers. ResNet50 addresses this issue by incorporating shortcut connections that bypass multiple layers, mitigating the vanishing gradient effect and facilitating the training of networks with hundreds of layers. As a result, ResNet50 has become a foundational model for

various visual recognition tasks, such as object detection, image classification, and, more recently, sign language recognition. Its ability to capture intricate features and hierarchical data makes it particularly effective for these applications. Figure 2 illustrates how ResNet50 can swiftly and accurately recognize sign language gestures. This architectural advancement has significantly impacted computer vision, providing a robust base for developing sophisticated models for recognizing American Sign Language (ASL).

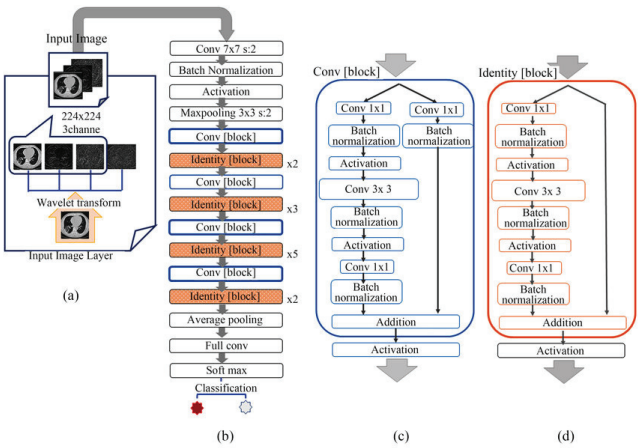

Fig. 2. Architecture of RestNet50

II. LITERATURE REVIEW

Sign language recognition (SLR) is a vital research area dedicated to enhancing communication for the deaf and hard-of-hearing community. This field focuses on understanding and interpreting sign language—a visual-gestural mode of communication involving hand gestures, facial expressions, and body movements. Recent advancements in machine learning and computer vision have significantly boosted the capabilities of SLR systems, offering promising solutions to communication barriers. Desai et al. [1] have made a notable contribution to this field with their development of ASL Citizen, a dataset designed to refine the accuracy of sign language recognition algorithms. ASL Citizen, created through community contributions, offers a diverse array of sign language samples, which facilitates robust model training and evaluation. In a novel approach, Song and colleagues [2] integrate sign language recognition with wearable technology. Their innovative wearable electronic skin, made from organohydrogel, is capable of detecting sign language gestures even in challenging environments, enabling real-time communication across various settings. Zhu et al. [4] propose a multiscale temporal network for continuous sign language recognition. Their method, which focuses on capturing both short- and long-term temporal dynamics of sign language motions, enhances recognition accuracy by incorporating multiscale temporal information, particularly useful for continuous recognition tasks. Alyami and team [5] apply a transformer-based approach to recognize isolated Arabic Sign Language (ASL). Their work highlights the importance of accommodating language-specific features in SLR. The proposed model, tailored to the unique aspects of Arabic sign language, such as hand shapes and movements, demonstrates effective recognition performance. Alves et al. [7] explore the use of skeleton image representation to enhance the recognition of Brazilian Sign Language (Libras). By capturing critical spatial and temporal details through skeletal image representation, their approach significantly improves the accuracy of Libras gesture recognition. Hu and colleagues [8][9][10] focus on refining continuous sign language recognition through enhanced image models. Their research aims to adapt existing image-based models for practical applications, resulting in improved performance in continuous SLR tasks by adjusting model parameters and incorporating domain-specific knowledge. Akdag and Baykan [11][12] introduce a multi-stream method for isolated sign language recognition based on finger features extracted from pose data. By integrating multiple data sources, such as hand gestures and movements, their approach boosts recognition accuracy, especially for complex sign language gestures. Wadhawan and Kumar [13][16] provide comprehensive reviews of deep learning-based SLR systems, addressing both static and dynamic sign gestures. Their work offers valuable insights into the evolution of SLR techniques, highlighting advancements in deep learning architectures and dataset creation, thus contributing to the progression of SLR technology. Looking ahead, SLR research continues to evolve with efforts in cross-lingual adaptation, real-time recognition, and multimodal integration. Future research could explore new sensor technologies, gesture detection across diverse cultural contexts, and the development of everyday SLR applications. These advancements aim to enhance global communication inclusivity for sign language users.

III. INPUT DATASET

For our research on American Sign Language (ASL) recognition, we utilize a dataset comprising training, testing, and evaluation subsets. The data, sourced from Kaggle, is organized by associating images with their respective labels and resizing them to 64x64 pixels using a function named `load_images`. The stratify parameter ensures an even distribution of data across all labels during the train-test split. By segmenting the data into training and testing groups, we gain insights into its composition. Specifically, the testing set contains 12,071 images, while the training set includes 48,281 images representing 40 different symbols. Additionally, a separate evaluation set of 8,000 images is prepared for validating and assessing model performance. This dataset forms the foundation for training and evaluating our ASL recognition algorithms, as depicted in Fig. 3.

```
Total number of symbols:  40
Number of training images:  48281
Number of testing images:  12071
Number of evaluation images:  8000
```

Fig. 3. Input Dataset Utilized

IV. PROPOSED METHODOLOGY

The proposed methodology outlines a structured approach for developing accurate and efficient American Sign Language (ASL) recognition models. Initially, it involves loading the dataset from specific directories to create training, testing, and evaluation sets, which form the foundation for developing and assessing the ASL recognition models. After loading, data preprocessing is carried out to prepare it for model training. This includes one-hot encoding

the labels into binary vectors, which facilitates effective categorical information interpretation for the models. Proper data formatting and preparation ensure the data is ready for input into the neural network models. Subsequently, the models are initialized, with a particular focus on two prominent architectures: ResNet50 and VGG16. Each model contributes unique architectural and performance advantages to the ASL recognition system. VGG16 is known for its deep convolutional layers that capture intricate details in images, while ResNet50 addresses the vanishing gradient problem with its residual connections, allowing the training of deeper networks. After initializing the models, their architectures are thoroughly examined to understand their underlying structures and parameters, which is crucial for optimizing performance and interpreting evaluation results. Performance is evaluated using confusion matrices for both VGG16 and ResNet50, which provide a detailed account of classification accuracy for each ASL gesture. These matrices highlight potential misclassifications and areas for improvement. Analyzing these matrices helps in understanding the strengths and weaknesses of each model, guiding future optimization efforts. Overall, this approach offers a systematic framework for developing and evaluating ASL recognition models, aiming to enhance communication accessibility for sign language users. The goal is to leverage deep learning to create robust and reliable ASL recognition systems through comprehensive testing and analysis.

V. RESULTS

A. *Model Accuracy Comparison*

Our comparative analysis of deep learning models for American Sign Language (ASL) recognition focused on VGG16 and ResNet50 architectures. The VGG16-based model achieved a remarkable 99.992% accuracy on test images and a perfect 100% accuracy on the assessment dataset. Meanwhile, the ResNet50 model recorded 99.95% accuracy on test images and also 100% accuracy on the evaluation dataset. These results indicate that both models perform exceptionally well in recognizing ASL movements, though with minor differences. The high accuracy rates of VGG16 and ResNet50 suggest their potential in facilitating communication for sign language users, thereby enhancing accessibility and inclusivity. Future research into these models could further develop robust ASL recognition systems, leading to improved functionality and broader real-world applications, as illustrated in Figures 4 and 5.

```
Accuracy for test images: 99.992 %
Accuracy for evaluation images: 100.0 %
```

Fig. 4. Accuracy Of VGG16

```
Accuracy for test images: 99.95 %
Accuracy for evaluation images: 100.0 %
```

Fig. 5. Accuracy Of Resnet50

B. *VGG16 accuracy and loss plot*

The performance of the VGG16-based ASL recognition model is illustrated by its accuracy and loss graphs, which track its progress through various training epochs. The accuracy graph depicts how well the model performs on both training and testing datasets as training advances, with both datasets showing improved accuracy over time, indicating effective learning of ASL gesture recognition. The alignment of the accuracy curves for both datasets suggests that the model is capable of generalizing well to new, unseen data, a crucial factor for reliable ASL recognition. Meanwhile, the loss graph displays how the model's predictions deviate from the actual labels, with a downward trend in loss indicating fewer prediction errors and effective learning. The close match between training and testing loss curves suggests that the model is not overfitting, enhancing its reliability for real-world applications. Figures 6 and 7 demonstrate that the accuracy and loss metrics for the VGG16-based ASL recognition model effectively capture its training behavior and confirm its capability to recognize ASL gestures accurately.

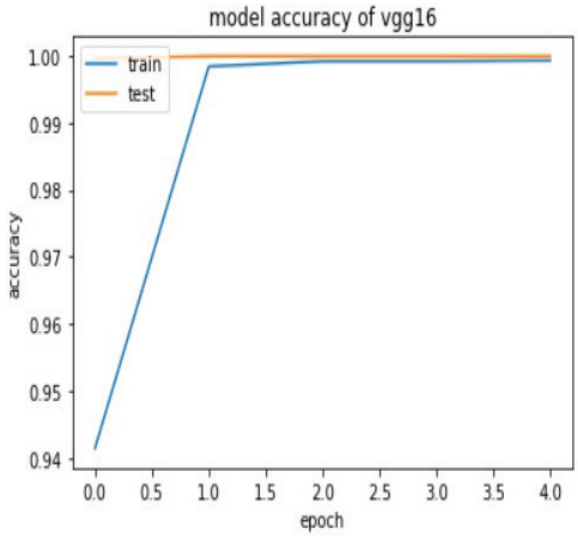

Fig. 6. Model accuracy of VGG16

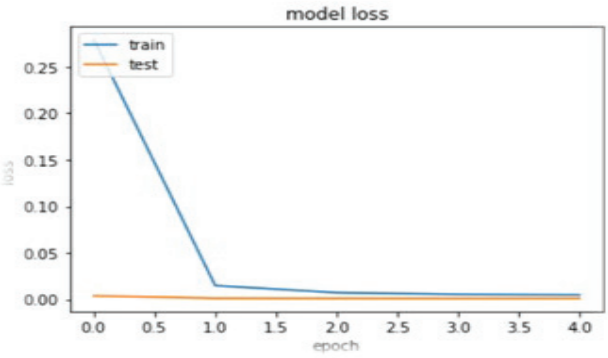

Fig. 7. Model loss

C. *Resnet50 accuracy and loss plot*

The graphs depicting accuracy and loss for the ResNet50-based ASL recognition model illustrate its performance throughout training and validation across multiple epochs. The accuracy graph tracks how well the model classifies ASL gestures within both the training and testing datasets over successive epochs. As training progresses, the accuracy for both datasets increases, highlighting the model's capability to learn and identify ASL signals effectively. The convergence of these curves suggests that the model generalizes well to new, unseen data,

demonstrating robustness in real-world applications. Conversely, the loss graph displays the model's prediction errors, represented by the gap between predicted and actual labels. A downward trend in loss indicates that the model is reducing errors in its predictions, confirming its learning effectiveness. The close alignment of training and testing loss curves further suggests that the model is avoiding overfitting and is thus more reliable for practical ASL recognition tasks. Overall, these accuracy and loss graphs provide insight into the ResNet50-based ASL model's training dynamics and its proficiency in accurately interpreting ASL gestures, as illustrated in Figures 8 and 19.

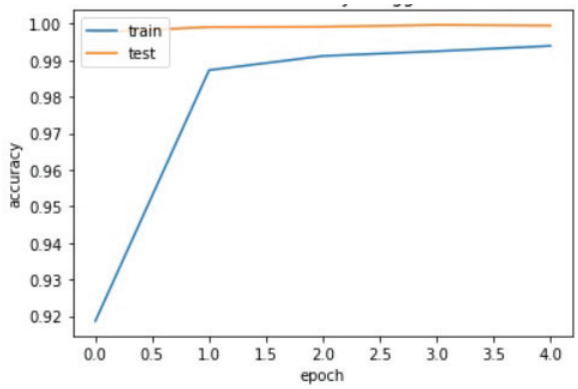

Fig. 8. Model Accuracy of RestNet.

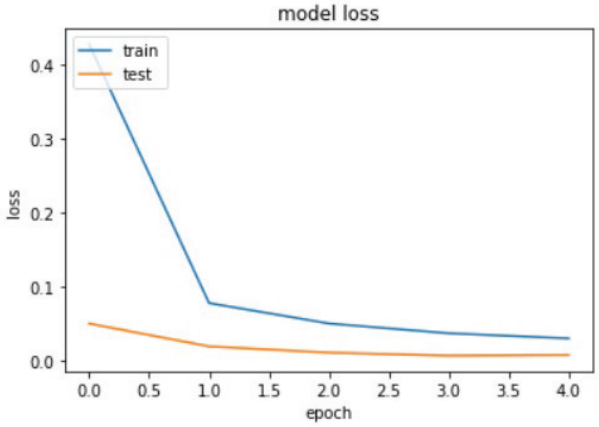

Fig. 9. Model Loss

D. *Single Image Prediction*

The image is first resized and converted into a NumPy array to match the input dimensions required by the model. It is then processed through the trained model to obtain predictions. For each ASL gesture, the model calculates the probability for each class, with the predicted class being the one with the highest probability. The result is mapped to the corresponding ASL gesture label using a series of conditional statements. The label with the highest probability is assigned to the gesture. Finally, the predicted label is displayed on the console. This process illustrates how the trained ASL recognition model is used to identify the ASL gesture represented in an individual image.

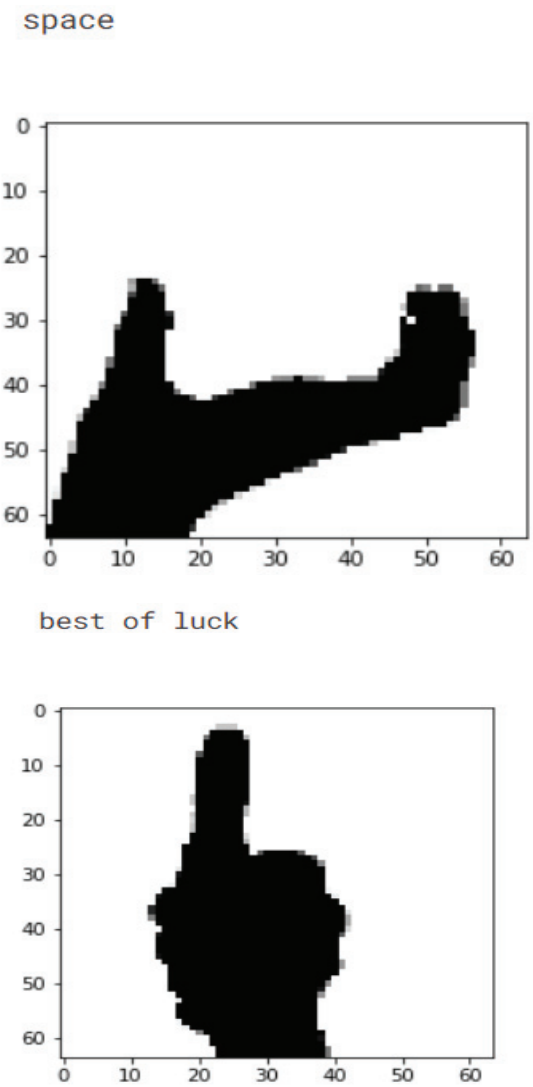

Fig. 10.

VI. Conclusion

In summary, our study finds that deep learning models, particularly VGG16 and ResNet50, excel in recognizing American Sign Language (ASL). Both models achieved impressive accuracy in interpreting ASL gestures, with VGG16 reaching a remarkable 99.992% and ResNet50 attaining 99.95% accuracy on test images. These results highlight the potential of deep learning architectures to accurately identify ASL movements, opening new possibilities for developing inclusive and accessible communication platforms. Additionally, the single image prediction functionality demonstrated the practical application of these models, enabling real-time detection of ASL gestures from individual images. Overall, our findings underscore significant progress in deep learning-based ASL recognition, with profound implications for enhancing communication accessibility and inclusivity for sign language users.

REFERENCES

[1] Desai, A., Berger, L., Minakov, F., Milano, N., Singh, C., Pumphrey, K., Ladner, R., Daumé III, H., Lu, A.X., Caselli, N. and Bragg, D., 2024. ASL Citizen: A Community-Sourced Dataset for Advancing Isolated Sign Language Recognition. Advances in Neural Information Processing Systems, 36.

[2] Song, B., Dai, X., Fan, X. and Gu, H., 2024. Wearable multifunctional organohydrogel-based electronic skin for sign language recognition under complex environments. Journal of Materials Science & Technology, 181, pp.91-103.

[3] Song, B., Dai, X., Fan, X. and Gu, H., 2024. Wearable multifunctional organohydrogel-based electronic skin for sign language recognition under complex environments. Journal of Materials Science & Technology, 181, pp.91-103.

[4] Zhu, Q., Li, J., Yuan, F. and Gan, Q., 2024. Multiscale temporal network for continuous sign language recognition. Journal of Electronic Imaging, 33(2), pp.023059-023059.

[5] Alyami, S., Luqman, H. and Hammoudeh, M., 2024. Isolated Arabic Sign Language recognition using a transformer-based model and landmark keypoints. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(1), pp.1-19.

[6] Alyami, S., Luqman, H. and Hammoudeh, M., 2024. Isolated Arabic Sign Language recognition using a transformer-based model and landmark keypoints. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(1), pp.1-19.

[7] Alves, C.E.G., Boldt, F.D.A. and Paixão, T.M., 2024. Enhancing Brazilian Sign Language Recognition through Skeleton Image Representation. arXiv preprint arXiv:2404.19148.

[8] Hu, L., Shi, T., Gao, L., Liu, Z. and Feng, W., 2024. Improving Continuous Sign Language Recognition with Adapted Image Models. arXiv preprint arXiv:2404.08226.

[9] Hu, L., Shi, T., Gao, L., Liu, Z. and Feng, W., 2024. Improving Continuous Sign Language Recognition with Adapted Image Models. arXiv preprint arXiv:2404.08226.

[10] Hu, L., Shi, T., Gao, L., Liu, Z. and Feng, W., 2024. Improving Continuous Sign Language Recognition with Adapted Image Models. arXiv preprint arXiv:2404.08226.

[11] Akdag, A. and Baykan, O.K., 2024. Multi-Stream Isolated Sign Language Recognition Based on Finger Features Derived from Pose Data. Electronics, 13(8), p.1591.

[12] Akdag, A. and Baykan, O.K., 2024. Multi-Stream Isolated Sign Language Recognition Based on Finger Features Derived from Pose Data. Electronics, 13(8), p.1591.

[13] Wadhawan, A. and Kumar, P., 2021. Sign language recognition systems: A decade systematic literature review. Archives of Computational Methods in Engineering, 28, pp.785-813.

[14] Camgoz, N.C., Koller, O., Hadfield, S. and Bowden, R., 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10023-10033).

[15] Gill, K.S., Anand, V., Gupta, R. and Pahwa, V., 2023, July. Insect Classification using Deep Convolutional Neural Networks and Transfer Learning On MobileNetV3 Model. In 2023 World Conference on Communication & Computing (WCONF) (pp. 1-5). IEEE.

[16] Wadhawan, A. and Kumar, P., 2020. Deep learning-based sign language recognition system for static signs. Neural computing and applications, 32(12), pp.7957-7968.

[17] Li, D., Rodriguez, C., Yu, X. and Li, H., 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 1459-1469).

[18] Gill, K.S., Sharma, A., Anand, V. and Gupta, R., 2023, March. Flower Classification Utilisizing Tensor Processing Unit Mechanism. In 2023 2nd International Conference for Innovation in Technology (INOCON) (pp. 1-5). IEEE.

[19] Min, Y., Hao, A., Chai, X. and Chen, X., 2021. Visual alignment constraint for continuous sign language recognition. In proceedings of the IEEE/CVF international conference on computer vision (pp. 11542-11551).

[20] Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K. and Fu, Y., 2021. Skeleton aware multi-modal sign language recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3413-3423).