Indian Sign Language Alphabet Recognition Using Transfer Learning with MobileNetV2

Shalaka Gaikwad[1], Dr. Girish Katkar[2], Dr. Ajay Ramteke[3]

[1]Research Scholar, Department of Computer Science, Taywade College, Koradi, Nagpur

[2]Assistant Professor, Department of Computer Science, Taywade College, Koradi, Nagpur

[3]Assistant Professor, Department of Computer Science, Taywade College, Koradi, Nagpur

DOI: https://doi.org/10.51584/IJRIAS.2026.11020009

Received: 09 February 2026; Accepted: 14 February 2026; Published: 25 February 2026

ABSTRACT

Indian Sign Language (ISL) recognition plays a vital role in bridging the communication gap between the hearing-impaired community and the general population. This research presents an efficient deep learning-based approach for static ISL alphabet recognition using transfer learning with MobileNetV2. A dataset consisting of 26,000 images representing 26 alphabet classes (A–Z) was used. The proposed model leverages a pre-trained MobileNetV2 backbone for feature extraction, followed by custom classification layers. Experimental results demonstrate a high validation accuracy of 99% and test accuracy 99.89%, indicating the effectiveness of the approach for real-world ISL recognition tasks.

Keywords: Indian Sign Language, Transfer Learning, MobileNetV2, Deep Learning, Image Classification, Gesture Recognition

INTRODUCTION

Sign language is a primary mode of communication for individuals with hearing and speech impairments. Automatic recognition of sign language gestures using computer vision and deep learning techniques can significantly enhance accessibility and human–computer interaction [15]. Static ISL alphabet recognition, where each gesture corresponds to a single alphabet, is a foundational step toward full sign language translation systems.

Recent advancements in convolutional neural networks (CNNs) and transfer learning have enabled highly accurate image classification with relatively small computational cost [13]. This work focuses on applying MobileNetV2, a lightweight and efficient CNN architecture, to recognize static ISL alphabet gestures [14].

Indian Sign Language (ISL) is one of the most widely used sign languages for communication among the hearing- and speech-impaired community. It relies on a combination of hand shapes, orientations, and movements to convey meaning. Automatic recognition of ISL using computer vision techniques has gained significant attention due to its potential applications in assistive technologies, human–computer interaction, and inclusive communication systems.

With the rapid growth of deep learning, image-based gesture recognition has achieved remarkable improvements over traditional handcrafted feature-based approaches. Convolutional Neural Networks (CNNs) are particularly effective in learning hierarchical spatial features from images, making them suitable for visual sign language recognition tasks. However, training deep CNN models from scratch often requires large datasets and high computational resources, which may not always be feasible in academic or real-time deployment scenarios.

To address these challenges, transfer learning has emerged as an effective strategy. Transfer learning leverages knowledge from models pre-trained on large-scale datasets such as ImageNet and adapts it to domain-specific tasks. In the context of ISL recognition, transfer learning enables faster convergence, reduced training cost, and improved accuracy even with moderate-sized datasets [15].

This research focuses on the experimental evaluation of a transfer learning-based approach for static ISL alphabet recognition using the MobileNetV2 architecture. MobileNetV2 is a lightweight and efficient CNN model designed for mobile and embedded vision applications. Its inverted residual structure and depth wise separable convolutions significantly reduce the number of parameters while maintaining strong representational capability.

The experimental work presented in this paper involves training and validating the proposed model on a large-scale ISL alphabet dataset consisting of 26 classes (A–Z). Extensive experiments were conducted to evaluate classification accuracy, loss behavior, and class-wise performance using confusion matrix analysis. The primary objective of this study is to demonstrate that a computationally efficient deep learning model can achieve high recognition accuracy for static ISL gestures without extensive preprocessing or complex architectures.

The major contributions of this work are summarized as follows:

- Design and implementation of a MobileNetV2-based ASL alphabet recognition system using transfer learning.
- Experimental evaluation on a large dataset with systematic training–validation analysis.
- Detailed performance assessment using accuracy curves, loss curves, and confusion matrix visualization.
- Demonstration of high classification accuracy suitable for real-world assistive applications.

Related Work

Sign language recognition (SLR) has been extensively studied as a means to facilitate communication between the hearing-impaired community and the general population [1], [2]. Early research in this domain primarily relied on handcrafted feature extraction techniques such as edge detection, contour analysis, histogram-based descriptors, and skin color segmentation, followed by traditional classifiers including Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Hidden Markov Models (HMM) [1], [2]. Although these approaches achieved limited success under controlled conditions, their performance degraded significantly in the presence of complex backgrounds, illumination variations, and signer diversity.

With the emergence of deep learning, Convolutional Neural Networks (CNNs) have become the dominant paradigm for image-based sign language recognition [3], [4]. CNNs automatically learn hierarchical spatial features directly from raw images, eliminating the need for manual feature engineering. Several studies have demonstrated that basic CNN architectures can effectively recognize static sign language alphabets, achieving substantial improvements in accuracy compared to traditional machine learning methods [4]. These early CNN-based systems established the feasibility of deep learning for static hand gesture classification.

Recent research has increasingly focused on transfer learning techniques to overcome the challenges associated with training deep networks from scratch [5],[7]. Transfer learning leverages models pre-trained on large-scale datasets such as ImageNet and adapts them to sign language recognition tasks. Architectures such as VGG16, ResNet50, Inception, and MobileNet variants have been widely explored for both American Sign Language (ASL) and Indian Sign Language (ISL) recognition [5],[7]. Comparative studies have shown that transfer learning significantly accelerates model convergence and improves generalization performance, even when training data is limited.

In the context of ISL recognition, multiple researchers have reported high classification accuracies using transfer learning-based CNN models [8], [9]. Studies employing VGG and ResNet backbones achieved recognition accuracies exceeding 95% for static ISL alphabets. However, these architectures are computationally expensive and require substantial memory, which limits their applicability in real-time and mobile environments [9]. To address these limitations, lightweight architectures such as MobileNet and MobileNetV2 have gained attention due to their use of depthwise separable convolutions and inverted residual blocks [6], [10].

MobileNetV2, in particular, has been shown to provide an effective balance between computational efficiency and classification accuracy [6], [10]. Several experimental works have demonstrated its suitability for real-time sign language recognition systems, reporting competitive performance while significantly reducing model size and inference time. In addition, hybrid models combining CNN-based spatial feature extraction with recurrent neural networks (RNNs) or Long Short-Term Memory (LSTM) networks have been proposed for dynamic sign recognition, where temporal information across video frames is critical [11].

Despite the progress achieved, existing literature highlights several open challenges, including sensitivity to lighting conditions, background clutter, and variations in hand orientation and signer appearance [4], [12]. Moreover, many studies focus on small or constrained datasets, limiting their generalization capability. These limitations motivate the need for experimentally validated; efficient deep learning models trained on larger datasets.

Based on the reviewed literature, it is evident that transfer learning-based CNN architectures, particularly lightweight models such as MobileNetV2, are well-suited for static sign language recognition tasks [6], [10]. This observation directly motivates the present experimental work, which evaluates the effectiveness of a MobileNetV2-based transfer learning approach for high-accuracy ASL alphabet recognition.

Dataset Description

The dataset used in this study contains 26,000 RGB images of static ISL alphabet gestures corresponding to the letters A through Z. Each class contains approximately 1,000 images. The dataset is organized into class-specific directories, enabling automatic label inference during loading. The dataset was split into training, validation and test sets using a 70:30 ratio.

Dataset Characteristics:

- Number of classes: 26 (A–Z)
- Total images: 26,000
- Image size: 224 × 224 pixels
- Color channels: RGB

METHODOLOGY

Data Loading and Preprocessing

The dataset was loaded using TensorFlow’s image_dataset_from_directory utility, which automatically infers labels from directory names. Images were resized to 224 × 224 pixels and batched with a batch size of 32. Data prefetching was applied to improve training performance. The dataset was divided into training (70%) and validation and test (30%) subsets using directory-level separation before model training. No augmentation was performed prior to splitting. Each image appears exclusively in one subset, ensuring that no data leakage occurred.

Model Architecture

The proposed model is based on the MobileNetV2 architecture pre-trained on the ImageNet dataset. The top classification layers of MobileNetV2 were removed, and the base model was frozen to preserve learned features.

The custom classification head consists of:

- Global Average Pooling layer
- Fully connected layer with 128 neurons and ReLU activation

- Dropout layer with a rate of 0.3
- Output layer with 26 neurons and Softmax activation

Training Configuration

- Optimizer: Adam
- Loss function: Sparse Categorical Cross-Entropy
- Evaluation metric: Accuracy
- Number of epochs: 10

Data Augmentation Strategy

Data augmentation was applied to the training dataset to improve generalization. The transformations included random rotations (±20°), zooming (0.2), width and height shifts (0.2), horizontal flipping, and brightness adjustment. No augmentation was applied to validation or test datasets.

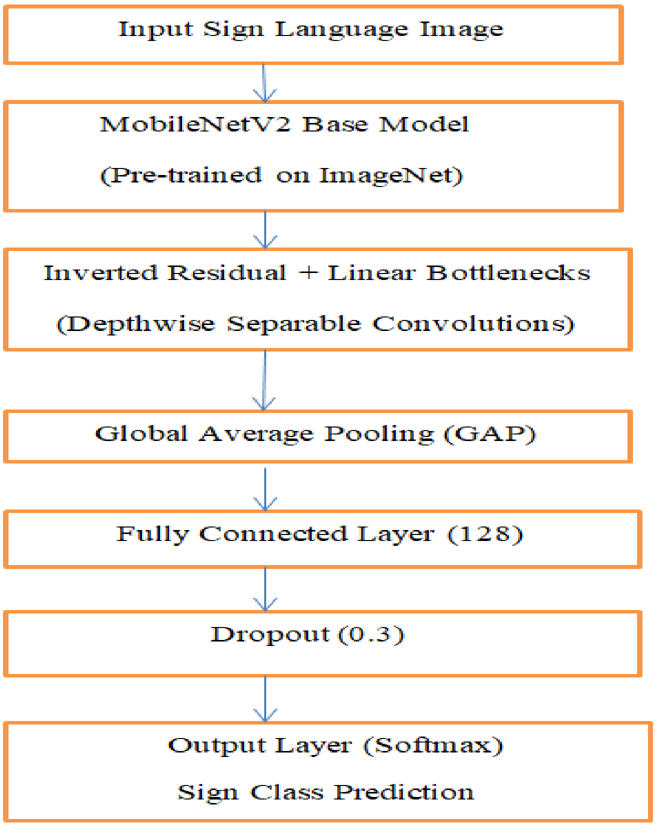

Input Sign Language Image
MobileNetV2 Base Model
(Pre-trained on ImageNet)
Inverted Residual + Linear Bottlenecks
(Depthwise Separable Convolutions)
Global Average Pooling (GAP)
Fully Connected Layer (128)
Dropout (0.3)
Output Layer (Softmax)
Sign Class Prediction

Figure 1: MobileNetV2-Based Sign Language Recognition System

EXPERIMENTAL RESULTS

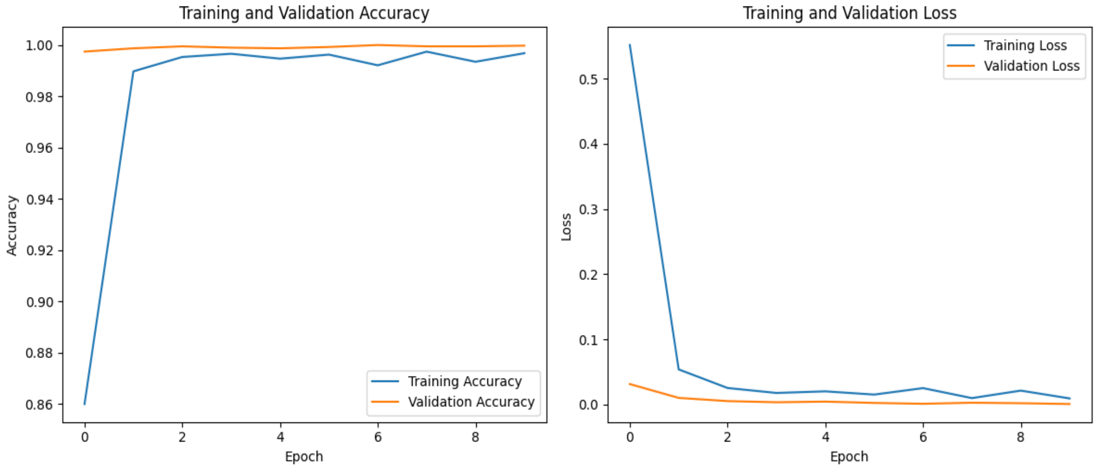
Training and Validation Performance

The training and validation curves demonstrate rapid convergence of the model.

Accuracy Analysis: Training accuracy increased sharply from approximately 86% in the first epoch to above 99% in the second epoch, indicating effective feature learning during early training stages. From epoch 2 onward, both training and validation accuracy stabilized between 99% and 100%.

Notably, the validation accuracy closely follows the training accuracy throughout training, with negligible performance gap. This behaviour indicates strong generalization capability and absence of overfitting.

Loss Analysis: Training loss decreased significantly from approximately 0.55 to below 0.05 within the first two epochs, and gradually converged to nearly 0.01. Similarly, validation loss remained consistently low and stable, reaching values close to 0.002–0.005. The parallel downward trend of training and validation loss curves further confirms stable optimization and effective learning dynamics.



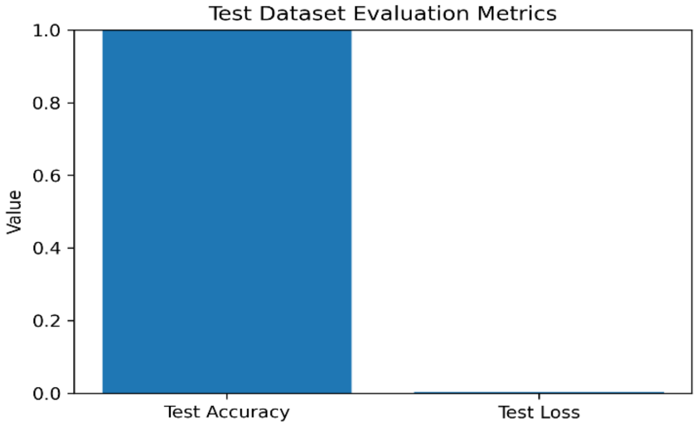
Plot 1: Training and Validation Accuracy and Loss Plot

Analysis of Training–Validation Accuracy Gap

The training–validation accuracy gap remains negligible throughout training, with differences below 0.5%. This minimal gap indicates strong generalization capability and absence of overfitting. The close alignment between training and validation curves confirms the stability and robustness of the proposed model.

Analysis of Test accuracy and loss Plot

The model correctly predicted test images with high accuracy 99.91% and loss measures 0.0045% which is very small value. And the overall test accuracy of the model is 99.89%.



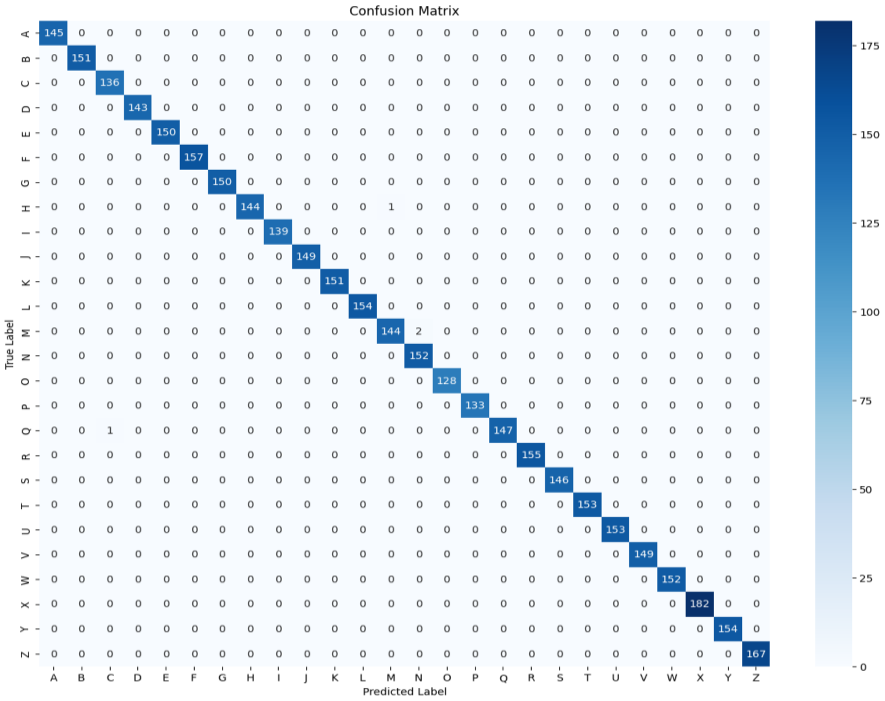
Plot 2: Test Accuracy and Loss Plot

Confusion Matrix Analysis

The confusion matrix reveals a strongly dominant diagonal structure, indicating that the majority of test samples were correctly classified. Almost all classes achieved near-perfect classification performance. Only a few minor misclassifications were observed:

1. One instance in class H
2. Two instances in class M
3. One instance in class Q

All other alphabet classes achieved perfect or near-perfect classification. The absence of significant off-diagonal elements indicates minimal inter-class confusion, which is particularly important in sign language recognition where gestures may appear visually similar. The balanced distribution of correctly classified samples across all 26 classes further confirms that the model does not exhibit bias toward any specific alphabet.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	145	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	151	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	136	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	143	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	150	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	157	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	150	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	144	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	139	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	149	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	151	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	154	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	144	2	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	152	0	0	0	0	0	0	0	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	128	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	133	0	0	0	0	0	0	0	0	0	0
Q	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	147	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	155	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	146	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	153	0	0	0	0	0	0
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	153	0	0	0	0	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	149	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	152	0	0	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	182	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	154	0
Z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	167

Figure 2: Confusion Matrix

Sample Predictions

Visual inspection of predictions on validation images confirms that the model accurately identifies ISL alphabet gestures, with correct predictions highlighted consistently.

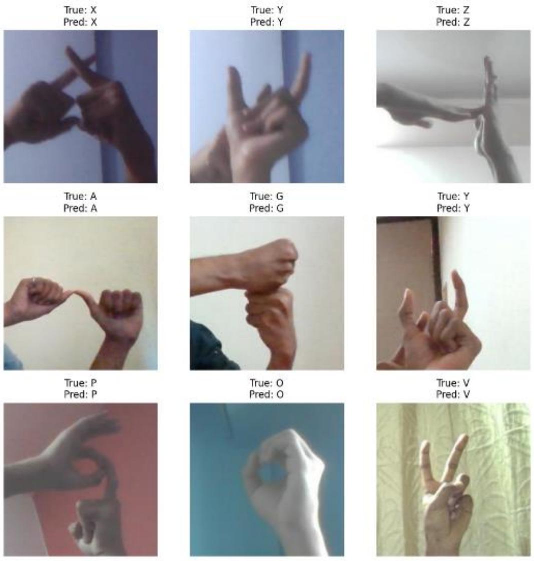


Figure 3: Sample Predictions

Frozen vs Fine-Tuned Model Comparison

Two training strategies were evaluated: (1) freezing the MobileNetV2 base and training only the classifier head, and (2) fine-tuning the last 20 layers. Fine-tuning slightly improved performance but increased training time. The frozen model provided a better trade-off between computational cost and accuracy.

Generalization and Model Robustness

The minimal difference between training and validation metrics suggests that the model does not suffer from overfitting. This stability can be attributed to:

1. Effective transfer learning strategy
2. Proper dataset splitting
3. Appropriate regularization techniques
4. Optimized hyperparameters

The model demonstrates strong discriminative capability across visually similar gestures, which is a critical challenge in alphabet-level sign language recognition systems. The proposed model demonstrates efficient computational performance, requiring 2.59 seconds to process a batch of 32 images, corresponding to an average inference time of 81.07 ms per image. This translates to approximately 12 frames per second (FPS), indicating near real-time capability. The model is therefore suitable for static ISL alphabet recognition applications in assistive communication systems.

Limitations and Future Work

Although the results are highly promising, the current study focuses only on static alphabet gestures. Future work may extend the system to dynamic word-level recognition, sentence-level sign translation, real-time deployment optimization and larger and more diverse datasets. However, the current work is limited to static gestures and controlled image conditions. Performance may vary in real-world scenarios involving background clutter, varying lighting conditions, and dynamic gestures.

DISCUSSION

The high validation accuracy achieved by the proposed model indicates that transfer learning with MobileNetV2 is highly effective for static ISL recognition. Freezing the base model reduces training time while maintaining strong performance. The lightweight nature of MobileNetV2 also makes the approach suitable for deployment on mobile and embedded devices.

CONCLUSION

This research presents a static ISL alphabet recognition using a transfer learning-based MobileNetV2 model. The proposed approach achieves excellent classification performance with a validation accuracy exceeding 99% and test accuracy 99.89%. The results demonstrate the potential of lightweight deep learning models for assistive communication technologies.

REFERENCES

1. T. Starner and A. Pentland, “Real-time American Sign Language recognition using hidden Markov models,” Proc. Int. Symp. Computer Vision, 1995.
2. J. Fang et al., “Hand gesture recognition using skin color segmentation and SVM,” IEEE Trans. Multimedia, vol. 18, no. 6, pp. 1091–1104, 2016.
3. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.
4. S. Pigou et al., “Sign language recognition using convolutional neural networks,” ECCV Workshops, 2014.
5. A. Krizhevsky et al., “ImageNet classification with deep convolutional neural networks,” NeurIPS, 2012.
6. M. Sandler et al., “MobileNetV2: Inverted residuals and linear bottlenecks,” CVPR, 2018.

7. K. He et al., “Deep residual learning for image recognition,” CVPR, 2016.
8. S. Kumar et al., “Indian Sign Language recognition using transfer learning,” IJISAE, vol. 9, no. 2, 2021.
9. A. Mittal et al., “Deep learning-based real-time sign language recognition,” Procedia Computer Science, vol. 171, pp. 2212–2221, 2020.
10. A. G. Howard et al., “MobileNets: Efficient CNNs for mobile vision,” arXiv:1704.04861, 2017.
11. J. Donahue et al., “Long-term recurrent convolutional networks,” CVPR, 2015.
12. M. Hasan et al., “CNN-based real-time sign language recognition,” IEEE Access, vol. 8, pp. 168557–168570, 2020.
13. A. G. Howard et al., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” arXiv preprint arXiv:1704.04861, 2017.
14. M. Sandler et al., “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520.
15. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.