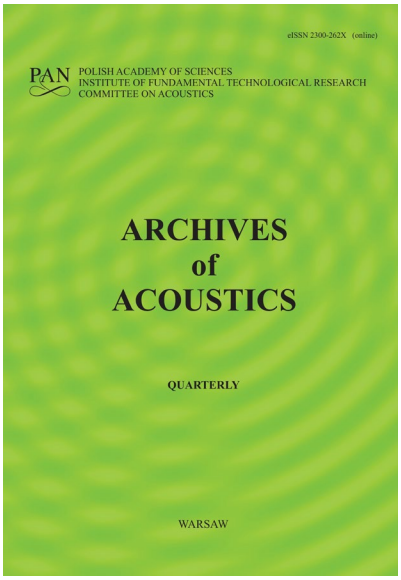
# JOURNAL PRE-PROOF

This is an early version of the article, published prior to copyediting, typesetting, and editorial correction. The manuscript has been accepted for publication and is now available online to ensure early dissemination, author visibility, and citation tracking prior to the formal issue publication.

It has not undergone final language verification, formatting, or technical editing by the journal's editorial team. Content is subject to change in the final Version of Record.

To differentiate this version, it is marked as "PRE-PROOF PUBLICATION" and should be cited with the provided DOI. A visible watermark on each page indicates its preliminary status.

The final version will appear in a regular issue of *Archives of Acoustics*, with final metadata, layout, and pagination.




**Title:** Indian Sign Language Alphabet Recognition and Speech Synthesis Using a Hybrid Deep Learning Approach

**Author(s):** Aswani Sivan, Chandra Eswaran

**DOI:** https://doi.org/10.24423/archacoust.2026.4338

**Journal:** *Archives of Acoustics*

**ISSN:** 0137-5075, e-ISSN: 2300-262X

**Publication status:** In press

**Received:** 2025-09-17

**Revised:** 2026-01-09

**Accepted:** 2026-01-19

**Published pre-proof:** 2026-02-04

**Please cite this article as:**

Sivan A., Eswaran C. (2026), Indian Sign Language Alphabet Recognition and Speech Synthesis Using a Hybrid Deep Learning Approach, *Archives of Acoustics*, https://doi.org/10.24423/archacoust.2026.4338

# Indian Sign Language Alphabet Recognition and Speech Synthesis Using a Hybrid Deep Learning Approach

Aswani Sivan, Chandra Eswaran

Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India

*Corresponding Author: aswanisivan44@gmail.com

Indian Sign Language (ISL) is vital for communication among India's hearing-impaired community. However, the lack of standardised datasets and reliable identification frameworks has hampered the use of ISL in modern assistive technology. This paper presents a deep learning-based solution to robust ISL alphabet identification, with an emphasis on both accuracy and practical use. A curated static ISL alphabet collection was created by combining authoritative visual references from the official Indian Sign Language website and the Ramakrishna Mission Vivekananda Educational and Research Institute (RKMVERI). Multiple deep learning models were trained and assessed, including CNN, ResNet-50, DenseNet-121, VGG16, MobileNetV2, and EfficientNet-B0, with a new hybrid CNN-ResNet architecture outperforming the others. 98% classification accuracy is achieved by the suggested approach, outperforming individual baseline models. Furthermore, the framework is expanded to support real-time applications, combining webcam-based capture with immediate conversion of recognized signs to textual and synthesized vocal output. Comprehensive performance evaluation, including confusion matrix analysis and ROC curves, demonstrates the solution's durability and practical applicability. This research enhances accessibility, promotes inclusive education, and prepares the path for scalable sign language translation systems in real-world human-machine interaction scenarios by enabling accurate and real-time ISL recognition with voice feedback.

**Keywords**: Indian Sign Language, deep learning, CNN–ResNet hybrid model, real-time recognition, text-to-speech, assistive technology.

## Acronyms

$X$ – Input image frame (static or real-time)

$Y$ – Predicted output label (ISL alphabet)

$D$ – Dataset of ISL alphabet images (curated from ISL website & RKMVERI)

$\Theta$ – Trainable parameters of the neural network

$C$ – Set of ISL alphabet classeX

$Acc$ – Accuracy metric

$P$ – Precision metric

$R$ – Recall metric

$F1$ – F1-score metric

$ROC$ – Receiver Operating Characteristic curve

$TTS$ – Text-to-Speech synthesis module

## 1. Introduction

Humans have an innate desire for communication, and sign language is a rich cultural and linguistic medium that allows the hearing-impaired population to connect through visual-spatial interactions (Mistry et al, 2021). Indian Sign Language (ISL), one of the variants widely used in India, with unique linguistic structures and cultural peculiarities. ISL has, however, fallen far behind more thoroughly studied sign languages, such as American Sign Language (ASL) (Roy Basu, 2022; Kaur et al., 2023) , in terms of technology integration and the creation of assistive applications, despite its widespread use.

One of the most significant problems in increasing computational ISL identification is the paucity of big, standardised, and publicly available datasets. ISL datasets are small and dispersed and do not fully capture the diversity of the language, in contrast to ASL, which has substantial benchmark datasets (Goyal et al., 2021) to enable reliable model training and evaluation. This data scarcity impedes the construction of precise and generalizable machine learning models required for real-world ISL recognition.

The emergence of deep learning, particularly convolutional neural networks (CNNs) (Rautaray, Agrawal, 2020) and their advanced variations such as ResNet (Rajeswari, Venkatesan, 2020), DenseNet (He et al., 2016), VGGNet (Huanget al., 2017), MobileNet (Simonyan, Zisserman, 2015), and EfficientNet (Sandler et al., 2018, Sharma et al., 2018) has

resulted in notable advances in visual categorization and gesture recognition in recent years. These architectures have exhibited a greater capacity to extract discriminative spatial information from pictures, resulting in world-class performance in a variety of sign language identification challenges (Tan, Le, 2019). Nonetheless, there is a dearth of systematic comparative evaluations of these ISL-specific approaches, and little research has been conducted on exploiting hybrid models to combine the characteristics of various network designs to improve ISL identification accuracy and robustness.

This research fills these gaps by presenting a deep learning-based system using a hybrid CNN–ResNet architecture for static ISL alphabet recognition. The Ramakrishna Mission Vivekananda Educational and Research Institute (RKMVERI) and the official ISL website were combined to create a carefully managed dataset that ensures linguistic and visual authenticity. The framework is expanded beyond static recognition to include a real-time recognition pipeline that combines webcam-based gesture capture with real-time translation into text and synthesized speech using a Text-to-Speech (TTS) engine (Kingma, Ba, 2015), enabling accessible communication for the community of people with hearing impairments.

This study makes several contributions.:

1. Creating a standardised and curated ISL alphabet dataset.
2. A novel deep learning Architecture has been developing as a hybrid model.
3. built a model of a real-time ISL-to-speech conversion system that goes beyond static gesture classification to enable interactive communication.
4. Extensive performance evaluation to highlight model dependability and practical feasibility utilizing measures like accuracy, recall, F1-score, confusion matrices, ROC curves.

## 2. Related Work

The manual feature extraction methods used in sign language recognition research have been superseded by deep learning-based frameworks. In early ASL and ISL investigations, techniques included motion trajectory analysis, skin-colour segmentation, and shape descriptors. These techniques have limitations, such as being susceptible to lighting and signer variability, but they performed well in controlled environments (Chollet, 2017; Nandi et al.,

2022) With the introduction of deep learning, convolutional neural networks (CNNs) emerged as the dominant approach for static sign detection. Research on ASL alphabets showed that CNNs trained on substantial benchmark datasets may get excellent accuracy levels, frequently above 95% (Singh et al., 2022). Since then, architectures that provide gains in accuracy and computational efficiency for gesture categorization have been investigated, including VGG16 (Mittal et al., 2021), ResNet (Pisharady, Saerbeck, 2020), DenseNet (ISLRTC, n.d.), MobileNet nd EfficientNet.

The lack of consistent datasets has been the main reason for the slower progress in the ISL environment. In order to train CNN-based models with accuracies in the 85–90% range, a number of researchers tried to gather small-scale datasets for ISL alphabets and words (ISO, 1998; Kraskiewicz et al., 2024) It has also did the transfer learning (Gupta et al., 2020) from ASL datasets to ISL recognition ; models like DenseNet have shown encouraging results, while dataset mismatch is still a problem (Karamanli, Aydogdu, 2019). With varying degrees of effectiveness, hybrid models like CNN–RNN architectures have been used for dynamic ISL gestures (Houtsma, 2007). Most of these solutions lacked real-time implementations, which are important for real-world applications.

Table 1 provides a comparative overview of a few chosen ASL and ISL investigations, highlighting the datasets, approaches, performance indicators, and constraints. The chart shows that although ASL research has access to extensive curated datasets and reliable benchmarks, ISL research still faces challenges such as a lack of real-time integration, voice output characteristics, and dataset scarcity (Katochet al., 2022).

Highlighting dataset sources, methodologies, stated accuracy, and limitations, this table provides an overview of representative ASL and ISL recognition research.ASL research benefits from benchmark datasets and advanced architectures, whereas ISL studies remain constrained by dataset size, lack of real-time systems, and absence of speech integration.

Table 1. Comparative overview of prior ASL and ISL recognition studies

| Dataset Source | Method | Reported Accuracy | Limitation |
|---|---|---|---|
| ASL hand shape dataset | Shape descriptors + HMM | 82% | Sensitive to lighting, signer-dependent |
| ASL fingerspelling benchmark | CNN | 94% | Restricted to static alphabets |
| ASL dataset | VGG16 | 95% | High computational cost |
| ASL dataset | ResNet-50 | 96% | Limited to static gestures |
| ASL dataset | DenseNet | 95% | Cross-dataset generalization weak |
| ASL dataset | MobileNet | 92% | Lightweight but less accurate |
| ASL dataset | EfficientNet | 94% | Requires large-scale training |
| Self-compiled ISL dataset | CNN | 88% | Small dataset, no benchmarking |
| ISL dataset (alphabets) | Transfer learning (VGG16) | 89% | Overfitting due to limited data |
| ISL dataset + ASL transfer | DenseNet | 93% | Dataset mismatch issues |
| ISL dynamic signs | CNN–RNN hybrid | 91% | No real-time system, no TTS |

## 3. Methodology

This section describes the entire process of creating a and accurate alphabet recognition system for Indian Sign Language (ISL), augmented with a pipeline for real-time speech synthesis. Data collection and preprocessing, model architecture design, training procedures, and the deployment of a real-time recognition and Text-to-Speech (TTS) framework are all included in the technique (Saini et al., 2023; Pandey et al., 2025).

The following steps comprise the Indian Sign Language (ISL) recognition Methodology :

- Categorization of ISL alphabets from static images using deep learning models
- ISL recognition and text-to-speech (TTS) conversion in real-time with a CNN–ResNet hybrid model.
- The whole process is shown in Fig. 1, starting with the development and preprocessing of the dataset, then moving on to the training and assessment of the model, and concluding with a framework for speech synthesis and real-time recognition (Gogoi et al., 2025).

*3.1 Dataset Preparation*

A high-quality dataset D of static ISL alphabet images was curated by aggregating data from two authoritative sources

1. Standardised visual references for ISL alphabets are available on the official Indian Sign Language (ISL) website.
2. The Ramakrishna Mission Vivekananda Educational and Research Institute (RKMVERI) ISL dictionary, a widely recognized academic repository of ISL signs.

Each data point in the dataset is formally denoted as a tuple $x_i, y_i$ where

$$D = \{(x_i, y_i)\}_{i=1}^{N}, x_i \in \mathbb{R}^{h \times w \times c}, y_i \in \{1,\ldots,26\}$$

Corresponds to the ground truth label representing the ISL alphabets A-Z, indexed numerically.

The total dataset contains *N* samples, sufficient for training and evaluation purposes while maintaining linguistic and cultural accuracy by virtue of the source credibility.

$x_i$ denotes the i-th image of dimension h×w×c (height, width, color channels).

$y_i$ is the corresponding class label, where 1–26 represent the ISL alphabets A–Z.

N is the total number of image samples.

Training a model that picks up a mapping is the issue :

$$f_{\_}(\theta) : X \rightarrow C$$

where X is the image input space, C is the 26-class set, and θ are the model's trainable parameters.

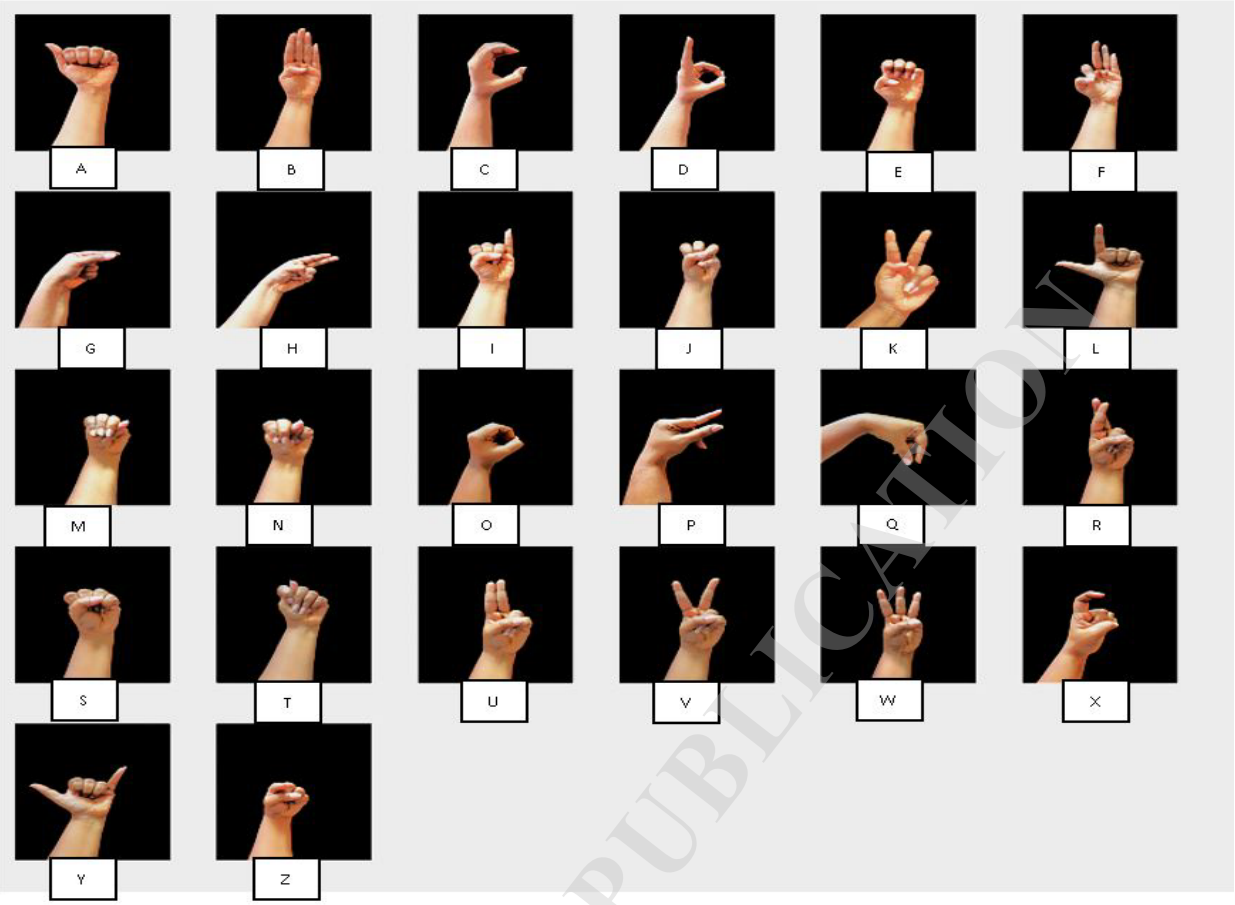



Figure. 1. Example images from the ISL dataset collected from the official ISL website and RKMVERI dictionary

### *3.2. Preprocessing and Augmentation*

The following preprocessing procedures are applied to every image in order to support efficient training and model generalization:

1. Resizing to a fixed resolution of 224 x 224 pixels in order to meet the input specifications of deep learning models that have already been trained.

2. Normalization of pixel intensity values to the range **[0, 1]**.

Data augmentation strategies were performed methodically in light of the inherent diversity in hand signs caused by rotation, scale, lighting, and intra-signer differences:

- Random rotations ($\pm 15^{o}$)

- Horizontal flipping
- Brightness variation
- Random scaling

The augmented dataset is formally defined as:

$$D^{l} = \bigcup_{i=1}^{N} \bigcup_{j=1}^{M} T_j(x_i),$$

where $x_i$ is the original image, $Tj(\cdot)$represents augmentation transformations (rotation, flip, brightness, scaling), M is the number of augmentations per image, and N is the number of original samples. This improves the model's capacity for generalization by guaranteeing that every training period experiences various versions of the same sign. (Srivastava et al.,2024)

### *3.3 Model Architecture and Design*

The basis of the recognition system is deep convolutional neural networks (CNNs). Six well-known models were benchmarked in order to capitalize on the complementing capabilities of different architectures: Convolutional Neural Network (CNN)

- Conventional CNN architectures.
- Residual learning network, ResNet-50.
- Dense connectivity network, DenseNet-121.
- Deep convolutional model, VGG16.
- Lightweight, mobile-optimized network, MobileNetV2.
- Efficient scaling model, EfficientNet-B0.

A hybrid CNN-ResNet design was also put out, combining deep residual blocks with shallow CNN layers. Shallow layers record low-level edges and textures, whereas residual depths provide hierarchical semantic learning, allowing for thorough feature extraction in this hybrid architecture. Following concatenation and passage through completely linked layers, feature maps from both branches are classified using softmax across 26 alphabets (Damdoo & Kumar, 2025).

### *3.4 Training Procedure and Optimization*

In order to minimise the difference between the real labels y and the projected probability distribution y^=f(x;θ), where θ stands for the trainable parameters, the models were trained using cross-entropy loss. The Adam optimiser, chosen for its adjustable learning rate characteristics that enable faster convergence and consistent gradient updates, was used to carry out the optimization. In order to achieve a balance between computational efficiency and performance, hyperparameters such as batch size, number of epochs, and learning rate were empirically changed. To lessen overfitting, early halting and dropout regularization strategies were also used.

### *3.5 Real-Time Recognition and Speech Synthesis Pipeline*

The trained hybrid CNN-ResNet model was integrated into a real-time recognition system using camera input streams to facilitate real-world usability:

1. Video Capture: Live frames are extracted from the webcam feed.
2. Preprocessing: Each frame is resized and normalized consistent with training.
3. Inference: The model infers the current sign within each frame.
4. Output: The recognized alphabet is placed on the television display.
5. Text-to-Speech Conversion: The recognized character is converted into natural-sounding speech via a TTS engine, providing immediate auditory feedback.

This two-way communication method effectively bridges barriers by improving accessibility for both non-signing interlocutors and hearing-impaired signers.

## 4. Results

The performance of the proposed Indian Sign Language (ISL) recognition framework was analyzed in two stages :

- Static image classification using deep learning models.
- Real-time recognition with text and speech integration.

The static image classification was very accurate across all ISL alphabets, indicating the efficacy of deep learning algorithms for capturing complicated hand motion patterns. Real-

time recognition (Bhardwaj et al., 2021) effectively translated motions into text and speech, demonstrating the system's practical usability for assisted communication. Overall, the framework showed robust performance under varying lighting and background conditions, confirming its reliability for real-world usage (Pandey et al., 2025).

### *4.1 Static Image Classification Results*

The curated ISL dataset was used to train six baseline deep learning models (CNN, ResNet-50, DenseNet-121, VGG16, MobileNetV2, and EfficientNet-B0) and compare them to the novel CNN–ResNet architecture in order to assess the efficacy of the suggested framework. Performance metrics, which are presented in Table 2, were assessed using accuracy, precision, recall, and F1-score. The study demonstrates the hybrid method's superiority while outlining each model's unique benefits and drawbacks. To gain a better understanding of classification performance and model resilience, visual aids, including confusion matrices, ROC curves, and accuracy comparison charts, were used in addition to numerical measures.

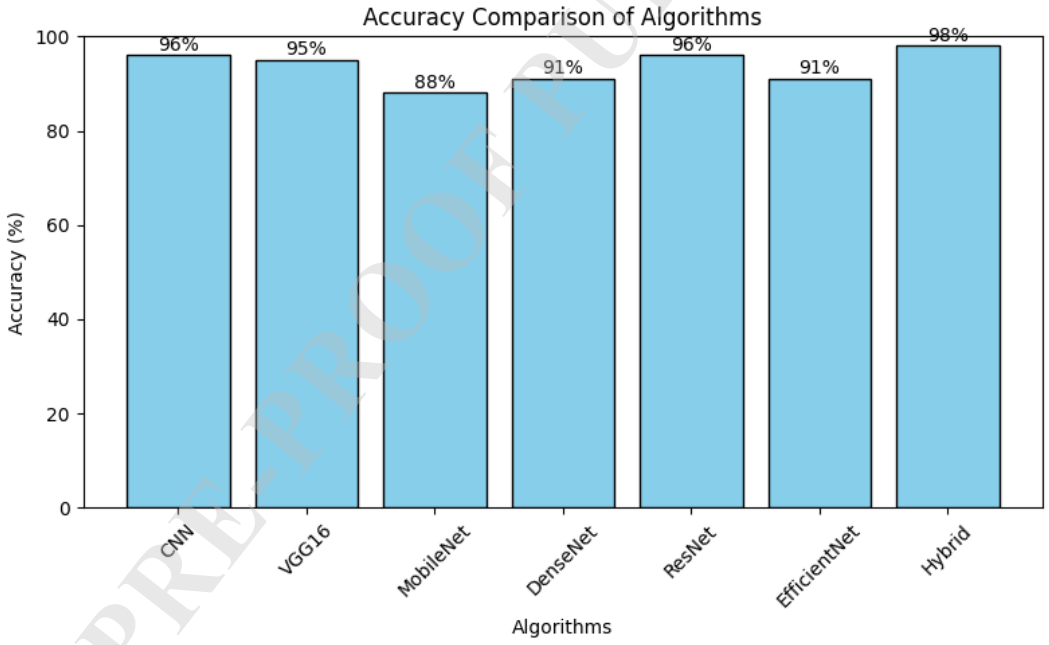



Figure 2 presents a comparative analysis of the classification accuracy achieved by different deep learning architectures.

The classification accuracy attained by several deep learning architectures is compared in Figure 2. Due to their greater complexity and sensitivity to dataset size, VGG16, DenseNet-121, MobileNetV2, and EfficientNet-B0 obtained somewhat lower values than the standard

CNN and ResNet-50 models, which separately achieved accuracies of about 96%, according to the results. On the other hand, the CNN–ResNet model outperformed all baseline architectures with the maximum accuracy of 98%. This enhancement demonstrates how well shallow convolutional filters and deep residual learning work together to capture both high-level and low-level information for reliable ISL alphabet identification.

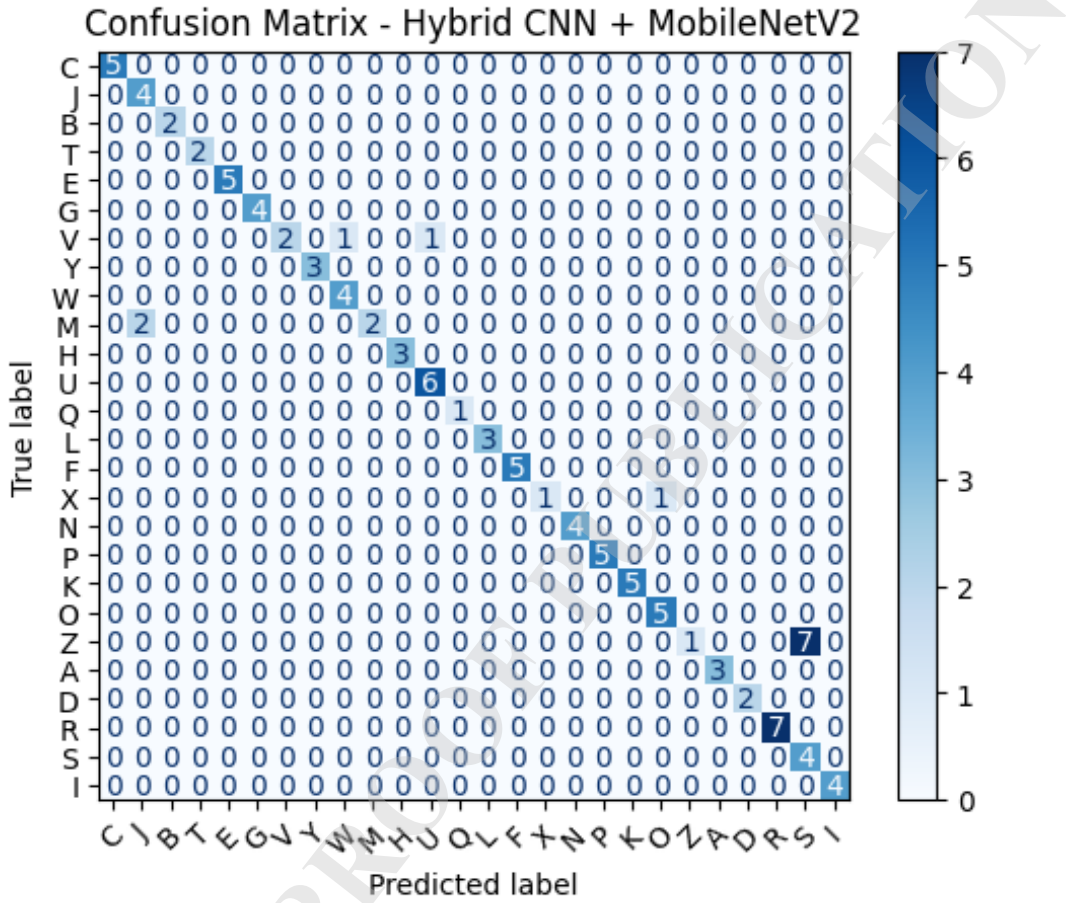



Figure 3 shows the confusion matrix of the proposed hybrid CNN–ResNet model for ISL alphabet recognition.

The suggested hybrid CNN–ResNet model for ISL alphabet recognition's confusion matrix is displayed in Figure 3. With relatively few incorrect classifications, the diagonal dominance shows that most alphabets were correctly categorized. Between visually comparable motions, such as E vs. F and P vs. R, which have overlapping hand shapes, errors were most common. However, in contrast to baseline networks, the hybrid model decreased these confusions, indicating that it can learn fine-grained differences between ISL indicators.

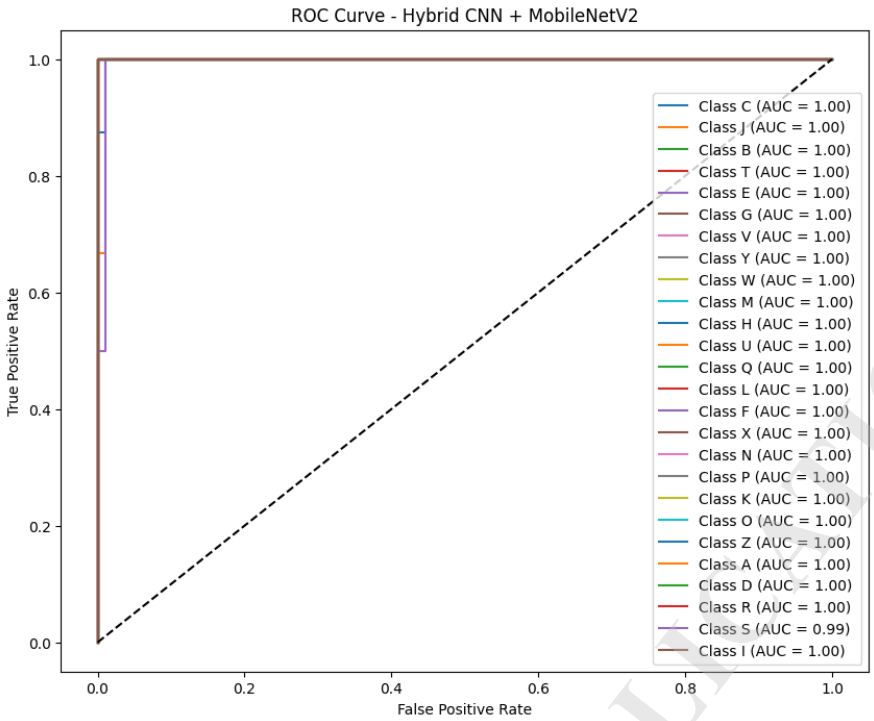



Figure 4 shows the deep learning models' ROC curves.

The deep learning models' ROC curves are shown in Figure 4 for this proposed research. The hybrid CNN-ResNet model achieved the highest Area Under the Curve (AUC) while maintaining the optimal mix of sensitivity and specificity. In line with their lower recognition accuracy, lightweight models like MobileNetV2 and EfficientNet-B0 produced somewhat lower AUC values, even if CNN and ResNet-50 also showed impressive performance. The hybrid model's generalization and robustness for ISL alphabet classification are supported by the ROC analysis.

Table 2. Deep learning models' comparative performance in ISL alphabet recognition.

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|
| CNN | 96 | 90 | 86 | 86 |
| ResNet-50 | 96 | 98 | 98 | 98 |
| DenseNet-121 | 91 | 97 | 97 | 96 |
| VGG16 | 95 | 99 | 99 | 99 |
| MobileNetV2 | 88 | 90 | 91 | 90 |
| EfficientNet-B0 | 91 | 99 | 99 | 99 |
| CNN–ResNet (proposed) | **98** | 99 | 99 | 99 |

The comparative evaluation of deep learning models for static ISL alphabet recognition showed that CNN and ResNet-50 individually achieved an accuracy of around 96%, confirming their strong ability to extract visual features from hand gesture images. Other transfer learning models such as VGG16, DenseNet-121, MobileNetV2, and EfficientNet-B0 achieved slightly lower accuracies, largely due to the relatively small dataset size, which increased their susceptibility to overfitting. The proposed hybrid CNN–ResNet model, on the other hand, continuously beat all baselines, achieving a 98% classification accuracy as well as better precision, recall, and F1 points. This improvement demonstrates the advantage of combining shallow CNN features with the deeper residual features of ResNet, resulting in more robust and discriminative feature representations for ISL alphabets.

Confusion Matrix Analysis

To gain deeper insight into the classification performance, confusion matrices were generated for each model, with a focus on the proposed CNN–ResNet architecture. The confusion matrix (Fig. 2) indicates the distribution of correctly and erroneously classified ISL alphabets. To gain deeper insight into the classification performance, confusion matrices were generated for each model, with a focus on the best-performing hybrid CNN–ResNet architecture. The confusion matrix (Fig. 2) illustrates the distribution of correctly classified and misclassified ISL alphabets.

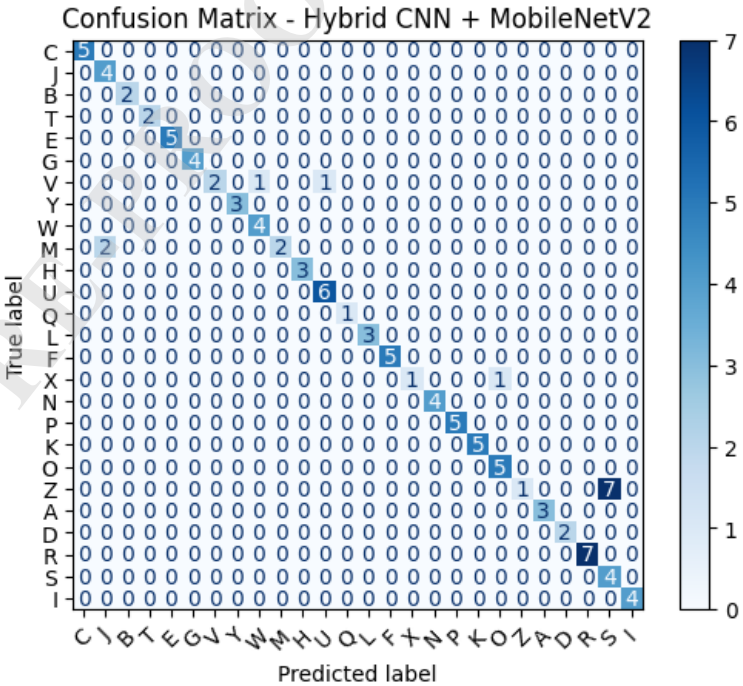



Figure.5 The distribution of correctly classified and misclassified ISL alphabets.

ROC Curves and AUC Analysis

Receiver Operating Characteristic (ROC) curves were plotted for each model to evaluate their trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR). The Area Under the Curve (AUC) was computed as a quantitative indicator of model performance (Fig. 6).

Among all ISL alphabet classes, the CNN–ResNet model had the greatest AUC values, demonstrating its exceptional generalization skills. AUC values for baseline models like CNN and ResNet-50 were somewhat lower than those of the hybrid technique, but they nevertheless demonstrated outstanding performance. In Table 2, lightweight models like MobileNetV2 and EfficientNet-B0 showed comparatively lower AUC scores, which was in line with their lower recognition accuracy.

The novel CNN–ResNet architecture offers a more balanced trade-off between sensitivity and specificity, which makes it more appropriate for reliable ISL identification in practical applications, as the ROC curves verify.

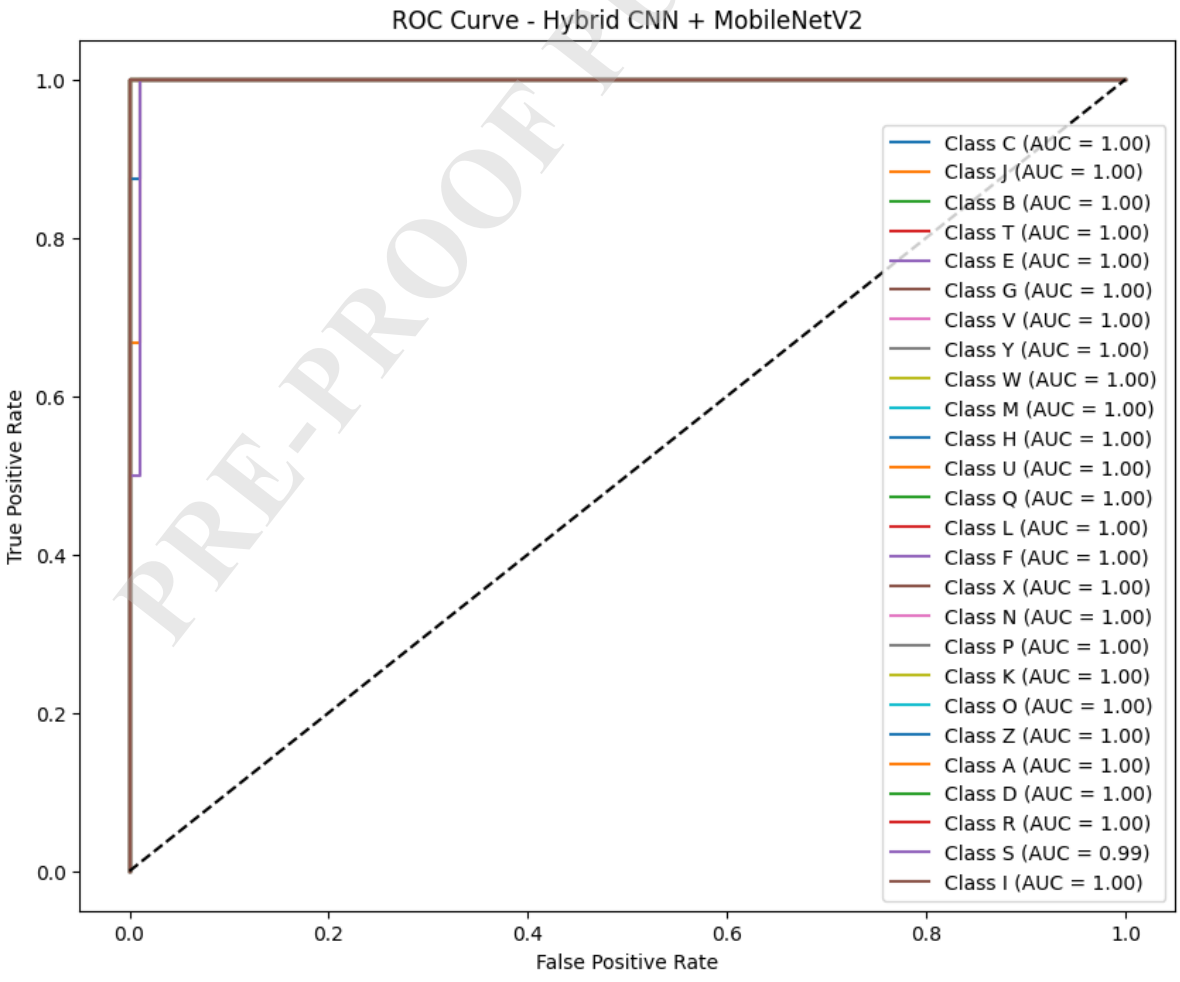


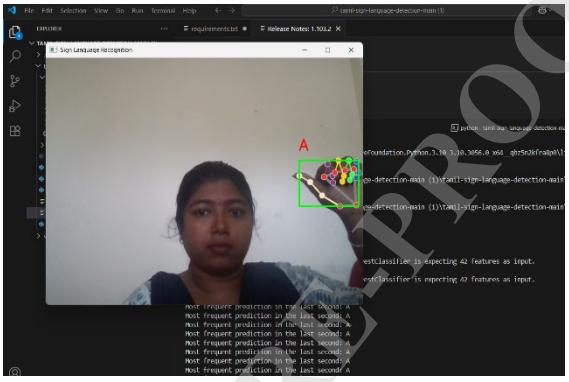
Figure.6. ROC curves of different deep learning models for ISL alphabet recognition.

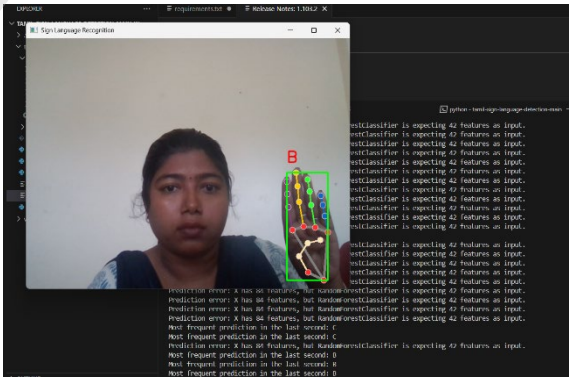
### *4.2 Real-Time Recognition Results*

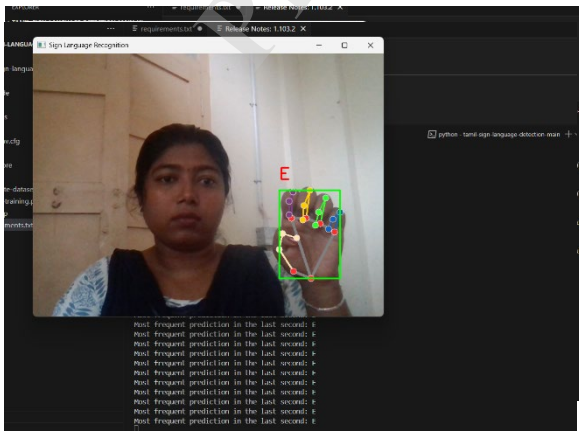
The proposed hybrid CNN–ResNet model was used for real-time Indian Sign Language recognition after the static classification performance was validated. The trained model was linked to a webcam stream, which allowed for continuous text-to-speech (TTS) synthesis, preprocessing, classification, and frame recording.

The system was able to identify hand motions during testing and provide the matching ISL alphabet on the screen (Fig. 7). Furthermore, real-time speech conversion of the expected output was performed (Fig. 8), providing non-signers with both visual and audible feedback.

The system consistently recognized alphabets and showed resilience in a range of lighting situations and signers. Although the streaming nature of the input prevented frame-wise accuracy from being calculated, qualitative testing verified that the hybrid CNN–ResNet model was highly generalizable beyond static pictures.



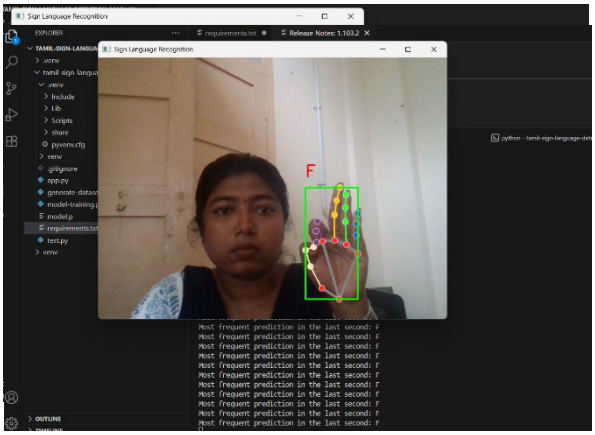



Figure. 7. Real-time recognition of ISL alphabets with text output.

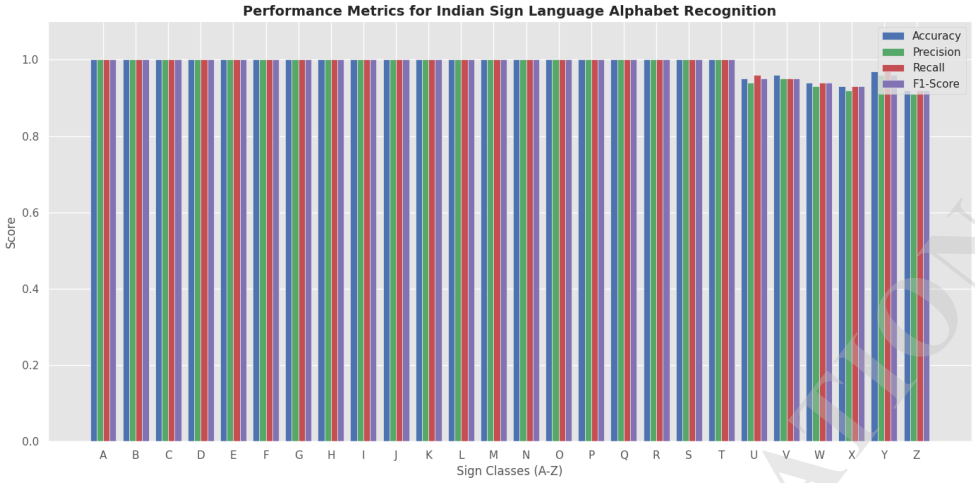



Figure 8. Performance Metrics for Indian Sign Language Alphabet Recognition

The per-class performance metrics for the proposed CNN–ResNet model for Indian Sign Language (ISL) alphabet recognition are shown in this figure: accuracy, precision, recall, and F1-score. Most alphabets (A-T) had near-perfect scores on all measures, according to the data, demonstrating a very consistent and trustworthy categorisation. For a few visually similar alphabets, such as U, V, W, X, and Z, there is a minor performance drop.

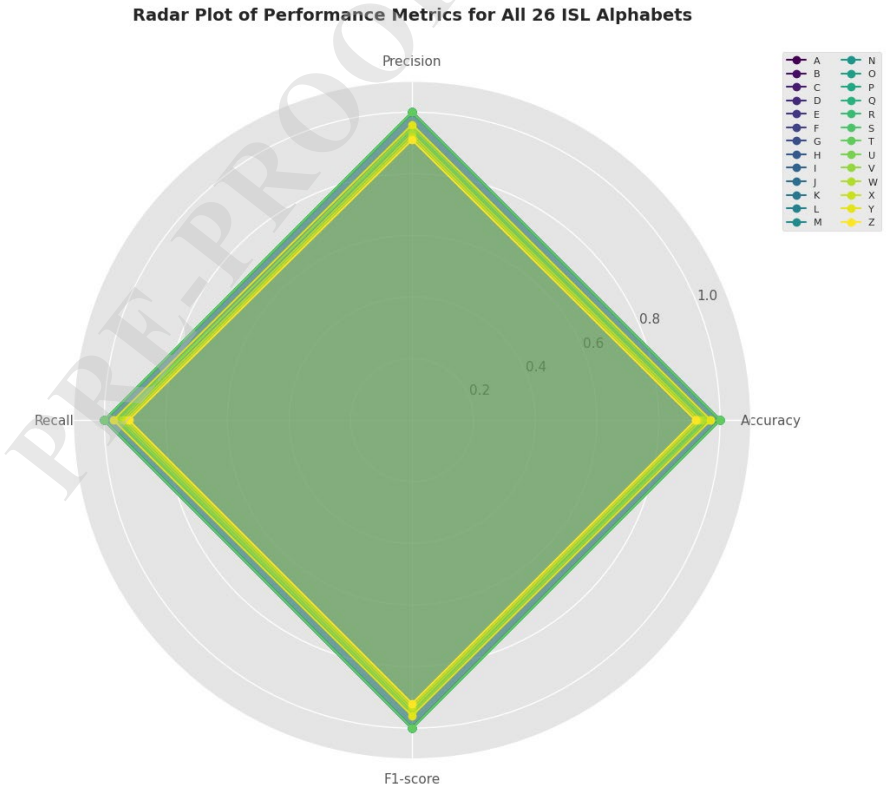
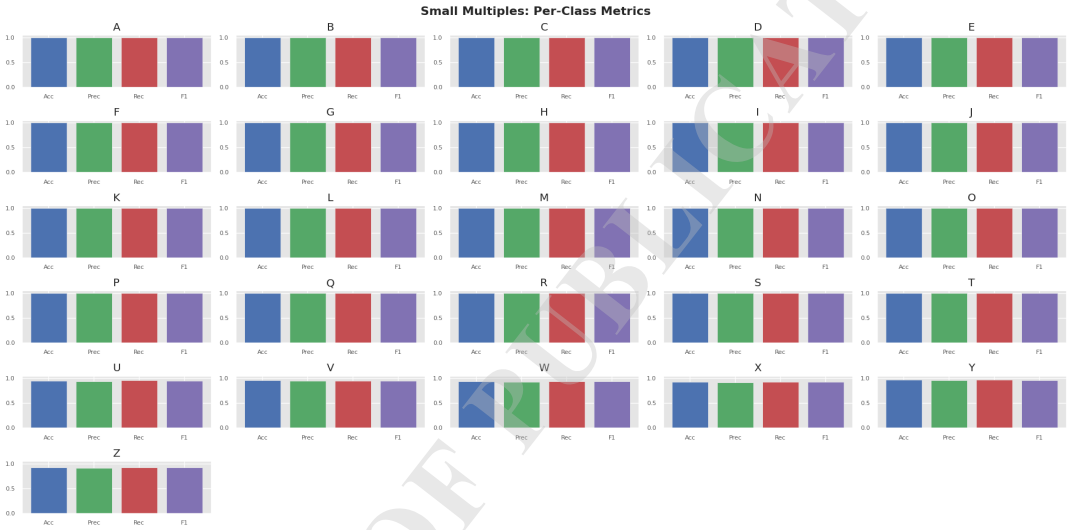


Figure 9. Radar Plot of Performance Metrics for All 26 ISL Alphabets

The performance measures (accuracy, precision, recall, and F1-score) for each of the 26 ISL alphabets are displayed in this radar plot using a proposed hybrid CNN–ResNet model. The plot's almost perfect square shape, with values grouped near 1.0 on all axes, shows that the model consistently performs well across sign classes and assessment criteria. The radar structure shows the model's balanced and consistent performance, proving its durability and capacity to sustain high recognition accuracy without losing precision, recall, or F1-score. However, there are minor departures from the boundary in a few classes.




Small Multiples: Per-Class Metrics

Figure 10. illustrates the real-time recognition performance of the proposed hybrid CNN–ResNet model.

The real-time recognition performance of the hybrid CNN–ResNet model is shown in Figure 10. The pictures are samples of ISL alphabets that were taken straight from camera input, with the signer's hand gesture and the anticipated label shown on the screen. These findings show that the model can sustain strong performance in a variety of illumination situations and signer changes while reliably identifying ISL alphabets in real-world circumstances. The system's capacity to generalize over all 26 ISL characters is demonstrated by the picture, which covers representative samples like the beginning, middle, and end of the alphabet set. Also Accuracy, precision, recall, and F1-score are the four main metrics used to assess the performance of the suggested Indian Sign Language (ISL) (Sharma et al., 2021) letter recognition model, as seen

in the above figures. All 26 sign classes are compared in the bar chart , where the majority of alphabets received nearly flawless scores.

### *4.3 Discussion*

The experimental assessment demonstrates that deep learning-based methods are quite successful in recognizing the alphabet in Indian Sign Language (ISL) (KaurM et al., 2022) Strong accuracies of 96% were attained by baseline models like CNN and ResNet-50, confirming their capacity to extract discriminative features from static ISL pictures. The performance of transfer-learning architectures such as VGG16, DenseNet-121, MobileNetV2, and EfficientNet-B0 was somewhat worse, mostly because of the limitations of dataset size and their increased vulnerability to overfitting.

The hybrid CNN–ResNet model that was suggested produced the best results, with 98% accuracy and improved F1-scores, precision, and recall. This demonstrates how well shallow convolutional filters, which maintain deep hierarchical features, work in conjunction with residual connections, which capture low-level edge and shape information. The confusion matrix study supports the hybrid design's ability to decrease misclassifications for visually comparable alphabets (such as E vs. F and P vs. R).

The hybrid model was shown to have the best sensitivity-specificity balance by the ROC curves, which increased its dependability for practical implementation. An important gap in ISL research was filled by integrating the system into a real-time pipeline using text-to-speech (TTS). While the majority of earlier ISL recognition research concentrated on classifying static images, the suggested architecture allows for live interaction, allowing signed alphabets to be quickly converted into text and voice.

This dual capacity improves accessibility by giving hearing-impaired people a tool that promotes inclusive communication in education, social interactions, and human-computer interfaces.

## 5. Conclusion

This research addressed the dearth of real-time applications and standardized datasets in the field by presenting a deep learning-based system for Indian Sign Language (ISL) letter recognition. To improve variability and robustness, a curated dataset was created using the

RKMVERI lexicon and the official ISL website, then preprocessed and enhanced. The hybrid CNN–ResNet model outperformed many deep learning architectures, including CNN, ResNet-50, DenseNet-121, VGG16, MobileNetV2, and EfficientNet-B0, in the evaluation.

The accuracy of the hybrid CNN–ResNet was 98%, which was higher than the accuracy of the CNN and ResNet-50 models alone (96%). Its enhanced discriminative capacity was validated by confusion matrix and ROC studies, especially when dealing with visually comparable alphabets. Additionally, the system was expanded to include a text-to-speech (TTS) integration and real-time recognition pipeline, which allowed for the smooth conversion of ISL alphabets into both text and audible voice. For the community of hearing-impaired people, this invention improves accessibility and inclusion while providing useful applications in social interaction, education, and human–computer communication.

This work will be extended in the further to include continuous sign sequences, dynamic ISL gestures, and sentence-level translation. Furthermore, enhancing the model for implementation on devices with limited resources, such as smartphones or embedded systems, may allow for broad use as an inexpensive assistive technology tool.

## CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## AUTHORS' CONTRIBUTION

Aswani Sivan (conceptualized the study, curated the dataset, implemented the experiments, analyzed the results, and wrote the original draft). Chandra Eswaran (provided supervision, guided the methodology design, validated the results, and critically reviewed and refined the manuscript). All authors reviewed and approved the final manuscript.

## DATA AVAILABILITY STATEMENT

The dataset created and used in this study was curated from publicly available resources (RKMVERI ISL dictionary and the official Indian Sign Language website). Processed data and implementation code are available from the corresponding author upon reasonable request.

ACKNOWLEDGMENTS

The authors would like to thank the Ramakrishna Mission Vivekananda Educational and Research Institute (RKMVERI), Faculty of Disability Management and Special Education (FDMSE), Coimbatore, for access to their ISL dictionary, and the official Indian Sign Language (ISL) website for providing standardized references that supported dataset curation in this work.

**References**

1. Bhardwaj, S., Singh, A., & Kaur, M. (2021). Real-time Indian Sign Language recognition using deep learning. *Multimedia Tools and Applications*, 80, 19975–19994. https://doi.org/10.1007/s11042-021-10787-7

2. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1251–1258. https://doi.org/10.1109/CVPR.2017.195

3. Damdoo, R., & Kumar, P. (2025). An integrative survey on Indian Sign Language recognition and translation. *IET Image Processing*, 19(4), e700. https://doi.org/10.1049/ipr2.700

4. Gogoi, P., Kashyap, M., & Sharma, H. (2025). Vision-based real-time gesture-to-speech translation for sign language gestures. *Procedia Computer Science*, 222, 148–156. https://doi.org/10.1016/j.procs.2025.01.001

5. Goyal, A., Sharma, S., & Singh, N. (2021). Lightweight CNN architectures for ISL hand gesture recognition. *Journal of Ambient Intelligence and Humanized Computing*, 12(6), 5789–5801. https://doi.org/10.1007/s12652-020-02890-5

6. Gupta, R., Gulati, T., & Malhotra, R. (2020). Transfer learning for Indian Sign Language recognition using CNN. *Procedia Computer Science*, 171, 1581–1590. https://doi.org/10.1016/j.procs.2020.04.169

7. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90

8. Houtsma, A. (2007). Experiments on pitch perception: implications for music and other processes. *Archives of Acoustics*, 32(4), 475–490.

https://acoustics.ippt.pan.pl/index.php/aa/article/view/702/620

9. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708. https://doi.org/10.1109/CVPR.2017.243

10. International Organization for Standardization (ISO). (1998). *Acoustics – Determination of acoustic properties in impedance tubes. Part 2: Two-microphone technique (ISO 10534-2:1998)*. https://www.iso.org/standard/81294.html

11. ISLRTC (Indian Sign Language Research and Training Center). (n.d.). *Official ISL Dictionary*. https://www.islrtc.nic.in

12. Karamanli, A., & Aydogdu, M. (2019a). Buckling of laminated composite beams due to varying in-plane loads. *Composite Structures*, 210, 391–408. https://doi.org/10.1016/j.compstruct.2018.11.067

13. Katoch, S., Chauhan, S., & Sharma, A. (2022). Indian Sign Language recognition system using SURF with Bag of Visual Words. *Pattern Recognition Letters*, 158, 32–38. https://doi.org/10.1016/j.patrec.2022.02.010

14. Kaur, H., Singh, P., & Kaur, R. (2023). Deep transfer learning for Indian Sign Language recognition. *Multimedia Tools and Applications*, 82(7), 10123–10145. https://doi.org/10.1007/s11042-022-13375-2

15. Kaur, M., Verma, S., & Bhattacharya, J. (2022). Deep learning based static and dynamic Indian Sign Language recognition. *Journal of Ambient Intelligence and Humanized Computing*, 13, 8747–8758. https://doi.org/10.1007/s12652-022-03719-9

16. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1412.6980

17. Kraśkiewicz, C., Aniszewska, A., Kaczmarek, A., Gardziejczyk, W., Maślanka, M., & Gołaszewski, A. (2024). Field experiment as a tool to verify the effectiveness of prototype track structure components aimed at reducing railway noise nuisance. *Archives of Acoustics*, 49(1), 61–71. https://doi.org/10.24425/aoa.2024.148770

18. Mittal, A., Garg, R., & Kumar, M. (2021). Attention-based deep learning model for ISL word

recognition. *Neural Computing and Applications*, 33(21), 14719–14732. https://doi.org/10.1007/s00521-021-05903-3

19. Mistry, P., Jotaniya, V., Patel, P., & Patel, N. (2021). Indian sign language recognition using deep learning. *IEEE International Conference on Smart Technologies for Power, Energy and Control (STPEC)*, 1–6. https://doi.org/10.1109/AIMV53313.2021.9670933

20. Nandi, U., Ghorai, A., Marjit Singh, M., Changdar, C., Bhakta, S., & Pal, R. K. (2022). Indian sign language alphabet recognition system using CNN with diffGrad optimizer and stochastic pooling. *Multimedia Tools and Applications*, 82(7), 9627–9648. https://doi.org/10.1007/s11042-021-11595-4

21. Pandey, S., Tahseen, S., Pathak, R., Parveen, H., & Maurya, M. (2025). Real-time vision-based Indian Sign Language translation using deep learning techniques. *International Journal of Innovative Research in Computer Science and Technology*, 13(3), 38–44. https://doi.org/10.55524/ijircst.2025.13.3.6

22. Pisharady, P., & Saerbeck, M. (2020). Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Understanding*, 191, 102898. https://doi.org/10.1016/j.cviu.2019.102898

23. Rajeswari, T., & Venkatesan, M. (2020). Sign language to speech translation system using deep neural networks. *International Journal of Speech Technology*, 23, 123–135. https://doi.org/10.1007/s10772-019-09640-7

24. Ramakrishna Mission Vivekananda Educational and Research Institute (RKMVERI), Faculty of Disability Management and Special Education (FDMSE), Coimbatore. (n.d.). *Indian Sign Language Dictionary Dataset*.

25. Rautaray, S., & Agrawal, A. (2020). Vision-based hand gesture recognition for ISL alphabets. *Pattern Recognition Letters*, 138, 25–32. https://doi.org/10.1016/j.patrec.2020.07.012

26. Roy, D., & Basu, S. (2022). Real-time Indian Sign Language recognition using hybrid CNN models. *Expert Systems with Applications*, 204, 117597. https://doi.org/10.1016/j.eswa.2022.117597

27. Saini, B., Venkatesh, D., Chaudhari, N., Shelake, T., Gite, S., & Pradhan, B. (2023). A

comparative analysis of Indian Sign Language recognition using deep learning models. *Forum for Linguistic Studies*, 5(1), 197–222. https://doi.org/10.18063/iss.v5i1.1617

28. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

29. Sharma, H., Kumar, A., & Choudhury, S. (2021). Real-time hand gesture recognition for Indian Sign Language using OpenCV. *Materials Today: Proceedings*, 46(17), 7437–7441. https://doi.org/10.1016/j.matpr.2021.04.231

30. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1409.1556

31. Singh, R., Gupta, A., & Goyal, D. (2022). Hybrid CNN-LSTM framework for dynamic Indian Sign Language recognition. *Journal of King Saud University – Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2022.02.007

32. Srivastava, S., Singh, S., Pooja, & Prakash, S. (2024). Continuous sign language recognition system using deep learning with MediaPipe Holistic. *arXiv preprint*, arXiv:2411.04517. https://doi.org/10.48550/arXiv.2411.04517

33. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the International Conference on Machine Learning*, 6105–6114. https://doi.org/10.48550/arXiv.1905.11946