iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: VI Month of publication: June 2025

DOI: https://doi.org/10.22214/ijraset.2025.72394

Sign Language Recognition with Deep Learning

Srushti Raskar[1], Neha Rane[2], Janhavi Kulkarni[3], Prof. N. D. Kuchekar[4]
Department of Electronics and Telecommunication Engineering, PVG's College of Engineering, Technology and Management, Pune

Abstract: Sign language serves as a vital means of communication for individuals who are deaf or speech-impaired. Despite its growing use, a communication barrier still exists between signers and non-signers. Recent advances in computer vision and deep learning have enabled the development of gesture recognition systems that can bridge this gap. In this research, we propose a real-time sign language recognition system that uses transfer learning with MobileNetV2 and a custom classification head. The system captures American Sign Language (ASL) gestures through a webcam and converts them into corresponding text in real time. The model is trained on a preprocessed ASL dataset and achieves high accuracy using efficient neural architectures, data augmentation techniques, and optimized training workflows. The system includes gamified learning levels—ranging from easy to complex—that provide feedback, scoring, and progress tracking to promote consistent user engagement and structured skill development.

Keywords: MediaPipe, MobileNetV2, Convolutional Neural Network, Image processing, sign language.

I. INTRODUCTION

People with impaired speech and hearing use sign language as a form of communication. Disabled People use these sign language gestures as a tool of non-verbal communication to express their emotions and thoughts to other common people. But these common people find it difficult to understand their expression; thus, trained sign language expertise is needed during medical and legal appointments and educational and training sessions. Over the past few years, there has been an increase in demand for these services. Other forms of services, such as video remote human interpretation using a high-speed Internet connection, have been introduced; thus, these services provide an easy-to-use sign language interpretation service that can be used and benefited from, yet they have major limitations. To address this, we use a custom CNN model to recognize gestures in sign language. We construct a convolutional neural network with 11 layers: four convolutional layers, three max-pooling layers, two dense layers, one flattening layer, and one dropout layer. We use a labeled American Sign Language (ASL) dataset with 26 classes (A–Z) to train the model to identify the gesture. The dataset contains the features of different augmented gestures. Introduced a custom CNN (Convolutional Neural Network) model to identify the sign from a video frame using OpenCV.

II. LITERATURE SURVEY

The model identifies American Sign Language using deep learning and computer vision [1]. The model gathers temporal and spatial attributes from video segments. Subsequently, to detect spatial attributes, we use Inception, a Convolutional Neural Network (CNN). Subsequently, to model temporal properties, we use a recurrent neural network (RNN). This research used an American Sign Language dataset.

A deep learning-based method for recognizing static sign language signals [2]. Humans can communicate proficiently using sign language, and significant research in computer vision is now under progress. The first research on Indian Sign Language (ISL) identification concentrated on the detection of significant, distinct hand signals; hence, only a limited number of ISL signs were selected for recognition. A cumulative collection of 35,000 signed photographs of over 100 static signs has been amassed from diverse users. The proposed system's efficacy is evaluated on around 50 CNN models.

Utilization of deep convolutional neural networks for the recognition of sign language [3]. This research utilizes a capture approach including continuous sign language video in selfie mode, enabling a hearing-impaired individual to freely use the SLR smartphone application. The dataset was developed with five unique people performing 200 signals from five different points of view under diverse background circumstances, addressing the scarcity of sign language datasets taken in smartphone selfies. In the video, each sign had around 60 frames. CNN training use three separate sample sizes, each including a varying number of persons and perspectives. The last two samples are used to evaluate the trained CNN.

Recognition of static sign language by deep learning [4]. The objective of the project was to develop a system capable of translating static sign language into its corresponding word counterparts, including letters, numerals, or fundamental static signals, to familiarize individuals with the fundamentals of sign language.

The researchers established an assessment technique and performed many tests to validate the significance of the system's non-signer functionalities. The solution received high marks for usability and learning efficacy throughout the evaluation.

A Comprehensive Study on Deep Learning-Based Methods for Sign Language Recognition [5]. A comparative experimental evaluation is conducted on computer vision-based sign language recognition systems. Recent deep neural network methodologies in this domain are used to conduct a comprehensive assessment of various publically accessible datasets. This project aims to enhance sign language recognition by mapping non-segmented video streams to glosses. This paper presents two novel sequence training criteria derived from the fields of voice and scene text recognition.

A deep learning-based method for recognizing static sign language signals [6]. The research discusses the use of deep learning via convolutional neural networks for the accurate identification of static signals in sign language recognition. This research amassed 35,000 photographs of signs, with each image depicting 100 static signs, contributed by diverse users. The proposed system's efficacy is evaluated on around 50 CNN models. Sign language is a complex and intricate system that relies on computer vision to interpret messages produced by hand motions with face expressions. It is a natural language used by those with hearing impairment to communicate. Sign language employs diverse hand gestures to convey letters, words, or sentences. We provide a pragmatic method for identifying ISL numbers, letters, and phrases in commonplace contexts. The suggested CNN design first employs convolutional layers, followed by ReLU and max-pooling layers.

Recognition of British Sign Language by Transfer Learning to American Sign Language using Late Fusion of Computer Vision and Leap Motion [7]. Researchers conducted many tests in both British and American Sign Language, emphasizing solitary sensory and multimodal methodologies. The results indicate that a multimodal approach surpasses the two individual sensors in training and classifying unknown inputs. This work included a preliminary transfer learning experiment from a substantial BSL dataset to a medium-sized ASL dataset, whereby the multimodality model was identified as the most effective for ASL classification after the transfer of weights from the BSL model. This research benchmarked and assessed all network topologies that were trained, compared, and ultimately fused to achieve multimodality for the first time. The precise classification of sign language, especially with unobserved data, enables autonomous completion of the process, offering a computerized method for interpreting non-spoken language in scenarios where interpretation is required yet inaccessible.

The mArSL Database and Pilot Study: Advancing Hybrid Multimodal Recognition of Manual and Non-Manual Arabic Sign Language [8]. A novel multi-modality ArSL dataset that integrates many modalities. The dataset consists of 6,748 video samples captured using Kinect V2 sensors, with fifty signs shown by four signers. This dataset will enable researchers to formulate and evaluate their approaches to advance the field. Furthermore, we used cutting-edge deep learning algorithms to examine the amalgamation of spatial and temporal attributes of various modalities, both manual and non-manual, for sign language recognition.

Recognition System for Thai Finger-Spelling Sign Language Using Deep Learning and Multi-Stroke Techniques [9]. A vision-based methodology was used to achieve semantic segmentation via dilated convolution for hand segmentation, optical flow separation for hand strokes, and feature learning and classification through a convolutional neural network (CNN). The five CNN architectures that dictate the forms were then compared. The initial format employed 64 filters, each measuring 3x3, across 7 layers; the subsequent format utilized 128 filters, also 3x3 in size, with 7 layers; the third format incrementally increased the number of filters while maintaining 7 layers, all featuring a uniform 3x3 filter size; the fourth format mirrored this structure; the final format was a structured configuration.

Utilizing k-Nearest Neighbors with Dynamic Time Warping and Convolutional Neural Network algorithms in wearable technology for sign language recognition [10]. The research includes a wearable electronics-based device for sign language recognition that utilizes two separate classification algorithms. The wearable electronics captured finger, wrist, and arm/forearm movements with a sensory glove and inertial measurement units. k-Nearest Neighbors using Dynamic Time Warping (a non-parametric methodology) and Convolutional Neural Networks used as classifiers (a parametric method) were implemented. Ten sign-words from Italian Sign Language were analyzed, including cose, grazie, and maestra, alongside globally recognized phrases such as google, internet, jogging, pizza, television, twitter, and ciao. Seven individuals, including five males and two females, aged 29 to 54 years, each replicated the signals 100 times (SD).

III. METHODOLOGY

The system begins by capturing real-time image frames using a webcam for gesture recognition. For each frame, MediaPipe Hands is employed to detect and extract 21 key landmarks of the hand. The identified hand region, or region of interest (ROI), is then resized to 224×224 pixels to meet the input requirements of the MobileNetV2 model. After resizing, the image is normalized so that pixel values fall within the range [0, 1], ensuring consistent input for the model.

The normalized image is passed through a pre-trained MobileNetV2 model (with frozen layers except for the last few), which extracts high-level visual features relevant to hand gestures. These features are further refined through custom convolutional neural network (CNN) layers, batch normalization, pooling, and dense layers designed to learn and specialize in American Sign Language (ASL) gesture patterns. Finally, a Dense layer with 26 output units, corresponding to the letters A–Z, uses softmax activation to predict the probability of each class. The class with the highest probability is selected as the recognized ASL letter.

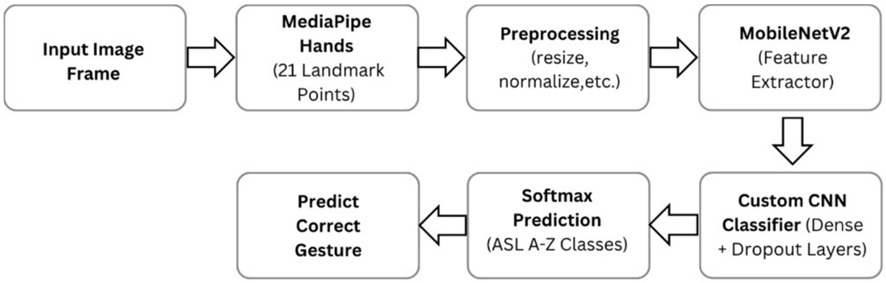

Input Image Frame
MediaPipe Hands (21 Landmark Points)
Preprocessing (resize, normalize,etc.)
MobileNetV2 (Feature Extractor)
Custom CNN Classifier (Dense + Dropout Layers)
Softmax Prediction (ASL A-Z Classes)
Predict Correct Gesture

Figure 1 : System Architecture

A. Dataset

The system is built for American Sign Language (ASL) gesture recognition using a convolutional neural network (CNN) approach. The dataset used consists of labeled hand gesture images organized into 26 categories, each representing a letter of the English alphabet (A–Z).

- Total Classes: 26 (corresponding to English alphabets A-Z)
- Training Images: 2,757
- Validation Images: 411
- Total Images: 3,168
- Image Format: RGB images loaded from directory structure
- Directory Structure: Each class is represented by a separate sub-folder under the main dataset directory.

B. Preprocessing

To standardize and augment the data before feeding it to the model, the following preprocessing steps were applied:

- Image Resizing: All images were resized to 224×224 pixels.
- Normalization: Pixel values were scaled to the [0, 1] range by dividing by 255.
- Data Augmentation: Implemented using ImageDataGenerator with the following parameters:
  - rescale : 1./255
  - rotation_range : 10
  - width_shift_range : 0.1
  - height_shift_range : 0.1
  - shear_range : 0.1
  - zoom_range : 0.1
  - fill_mode : 'nearest'

This data augmentation strategy increases the effective size and variability of the training set, reducing the risk of overfitting and improving generalization.

C. Model Architecture

The model leverages transfer learning using MobileNetV2, a lightweight and efficient deep neural network pre-trained on the ImageNet dataset.

D. Base Model

- Model: MobileNetV2
- Pre-trained Weights: ImageNet

- Top Layers: Removed (include_top=False)
- Trainable: Set to False during initial training to preserve learned features
- Input Shape: (224, 224, 3)

E. Custom Classification Head :

- GlobalAveragePooling2D() layer to reduce the feature maps
- Dropout(0.3) for regularization
- Dense(512, activation='relu') for deep feature learning
- Final layer: Dense(26, activation='softmax') to output class probabilities across 26 categories

This architecture provides a powerful combination of pre-trained feature extraction and custom classification tailored to ASL images.

F. Training Configuration

- Loss Function: Categorical Crossentropy (multi-class classification)
- Optimizer: Adam
- Epochs: 100
- Batch Size: 32
- Validation Split: Explicit validation folder (411 images)
- Framework: TensorFlow/Keras on Google Colab

IV. TECHNOLOGIES USED

1) TensorFlow / Keras: Used for building, training, and deploying the deep learning model. TensorFlow provides the computational backend while Keras offers a user-friendly interface for defining model layers and training configurations.
2) MobileNetV2: A lightweight and efficient deep learning architecture pre-trained on ImageNet. It is used here as a feature extractor via transfer learning to enhance accuracy with reduced computational cost, suitable for real-time ASL recognition.
3) OpenCV: Utilized for real-time video capture from the webcam and for preprocessing frames (resizing, grayscaling, etc.) before they are passed to the model.
4) MediaPipe: Used to detect and track hand landmarks in real-time. It helps isolate the region of interest (hand gestures), reducing background noise and improving model accuracy.
5) SQLite: A lightweight embedded database used to store user progress, accuracy reports, and session data locally. It enables tracking of individual practice results.
6) Python: The core language used for writing all parts of the system, from model training to real-time testing, with integration of AI, computer vision, and GUI/web frameworks.
7) Django: Used to build the backend of the ASL recognition web application, manage user authentication, scoring, and web routing.

V. FUTURE SCOPE

1) AI Integration for Subtitle-to-Sign Conversion: Real-time translation of video subtitles into sign language using AI-powered animated avatars.
2) Mobile App Integration: Cross-platform mobile applications for on-the-go sign language learning and communication. User-friendly interfaces to increase accessibility for diverse users.
3) Multi-Language and Dialect Support: Expansion to include multiple sign languages (e.g., ASL, BSL, ISL) and regional dialects. Customized signing styles for different linguistic and cultural communities.
4) Educational Platform Integration: Embedding sign language tools into e-learning platforms to aid Deaf and hard-of-hearing students.
5) Facial Expression and Emotion Recognition: Advanced detection of facial cues such as eyebrow movements, lip shapes, and emotional expressions. Emotion recognition to capture the signer’s tone and intent, enhancing the expressiveness of sign interpretation.

VI. RESULTS

The model was trained for 100 epochs with training and validation metrics recorded over time. Based on the final epoch:

- Training Accuracy: 97%
- Validation Accuracy: 92%
- Training Loss: 0.11
- Validation Loss: 0.20

These results show strong learning performance and generalization capability, supported by the use of data augmentation and a robust MobileNetV2 base.

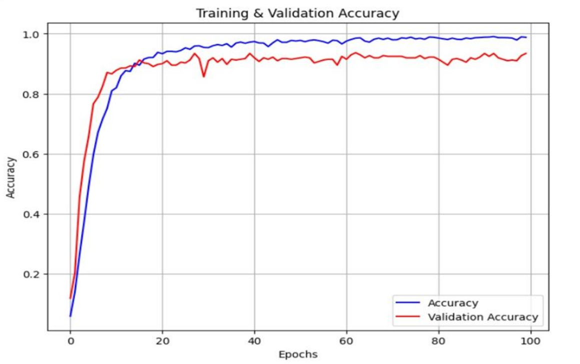


Figure 1.2 Training and Validation Accuracy

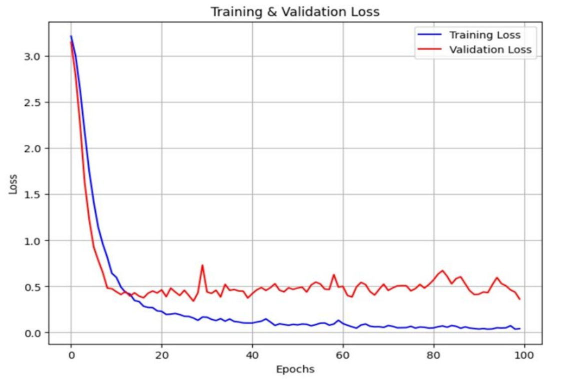
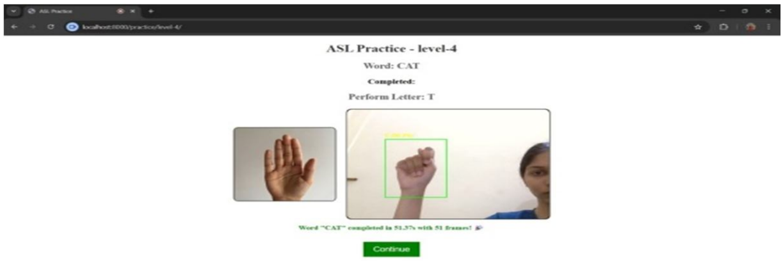


Figure 1.3 Training and Validation Loss

The graph below illustrates the classification of the system. The graphs illustrate the system's allocation of total inputs among several scenarios. The suggested methodology utilizes a mixture of CNNs that yields exceptional results universally. The input photographs were supplied for training, and assessments were conducted for performance evaluation using several classification models.

Figure 2 below illustrates a real-time assessment of the suggested model. The model demonstrates its capability to recognize sign language



ASL Practice - level-4

Word: CAT

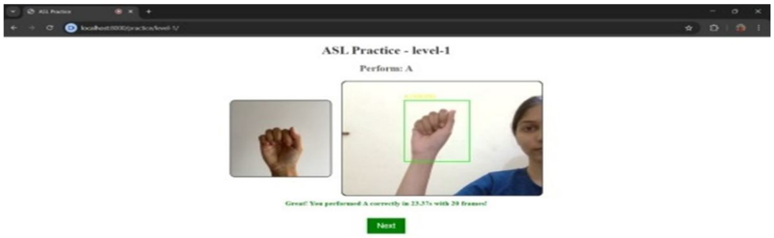
Completed:

Perform Letter: T

Word "CAT" completed in 51.37s with 51 frames! 🎉

Figure 2.1 : Sign Detection Level 1

This product is a Level 1 ASL Practice Module that uses a camera-based recognition system to assess the user's sign language motions in real time, offering feedback and accuracy ratings. It assists learners in practicing and enhancing their ASL fingerspelling skills.



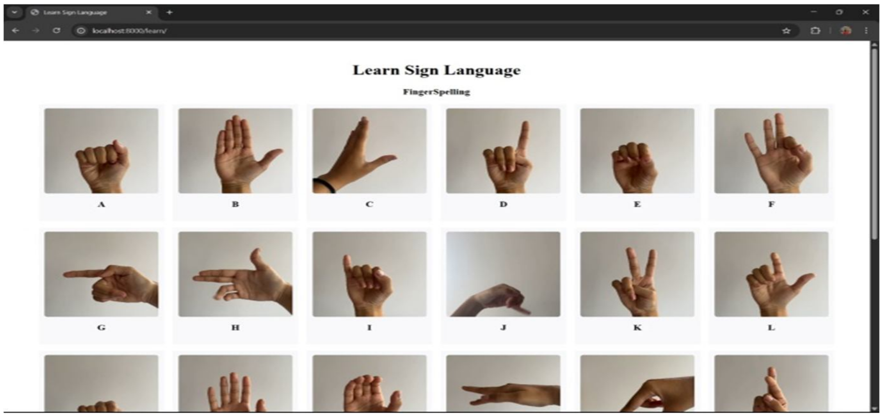
ASL Practice - level-1

Perform: A

Great! You performed A correctly in 23.37s with 20 frames!

Figure 2.2 : Sign Detection Level 4

The ASL Level-4 interface guides users to spell words like “CAT” using hand signs with real-time feedback and accuracy scoring.



Learn Sign Language

FingerSpelling

Figure 2.3 : Learn Section 1

The graphic depicts a web-based platform designed for learning American Sign Language (ASL) fingerspelling. It has a tidy, grid-format arrangement with hand sign pictures for the letters A through L, each accompanied by its caption underneath. The title of the page is "Learn Sign Language," and it operates locally. This program facilitates visual learning of the ASL alphabet for users.

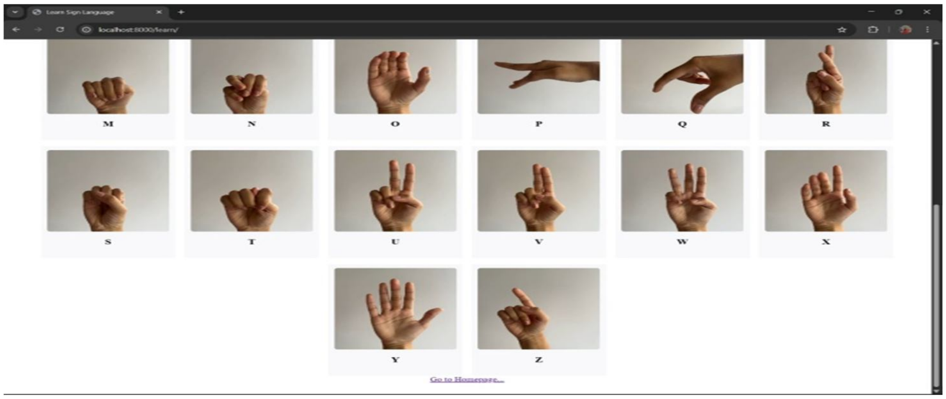


Figure 2.4 : Learn Section 2

VII. CONCLUSION

The proposed sign language recognition system leverages MediaPipe Hands for precise hand tracking and a MobileNet-based CNN model for accurate real-time gesture recognition by processing video input frame by frame, within an interactive web platform.Designed as a gamified learning environment, the system motivates users to practice and improve their sign language skills effectively.

The platform incorporates SQLite to manage user authentication and store performance scores securely. By fostering inclusive communication and breaking down language barriers, this system holds significant social relevance in promoting accessibility and understanding across diverse communities.

Overall, it offers an engaging and accessible solution to enhance communication and learning for a wide range of users.

REFERENCES

[1] M. Hafiz, A. Joshi, and A. Salinas, “Development of a Sign Language Recognition System Using Machine Learning,” Team E5 Final Report, pp. 1–60.

[2] H. Orovwode, O. Ibukun, and J. A. Abubakar, “A machine learning-driven web application for sign language learning,” Front. Artif. Intell., vol. 7, no. 1297347, pp. 1–10, Jun. 2024, doi: 10.3389/frai.2024.1297347.

[3] J. J. Raval and R. Gajjar, “Real-time Sign Language Recognition using Computer Vision,” in Proc. 2021 3rd Int. Conf. Signal Process. Commun. (ICSPC), Coimbatore, India, 2021, pp. 542–546, doi: 10.1109/ICSPC51351.2021.9451709.

[4] R. G. Rajan and M. J. Leo, “American Sign Language Alphabets Recognition using Hand Crafted and Deep Learning Features,” in Proc. 2020 Int. Conf. Inventive Comput. Technol. (ICICT), 2020.

[5] A. Wadhawan and P. Kumar, “Deep learning-based sign language recognition system for static signs,” Neural Comput. Appl., vol. 32, no. 12, pp. 7957–7968, 2020.

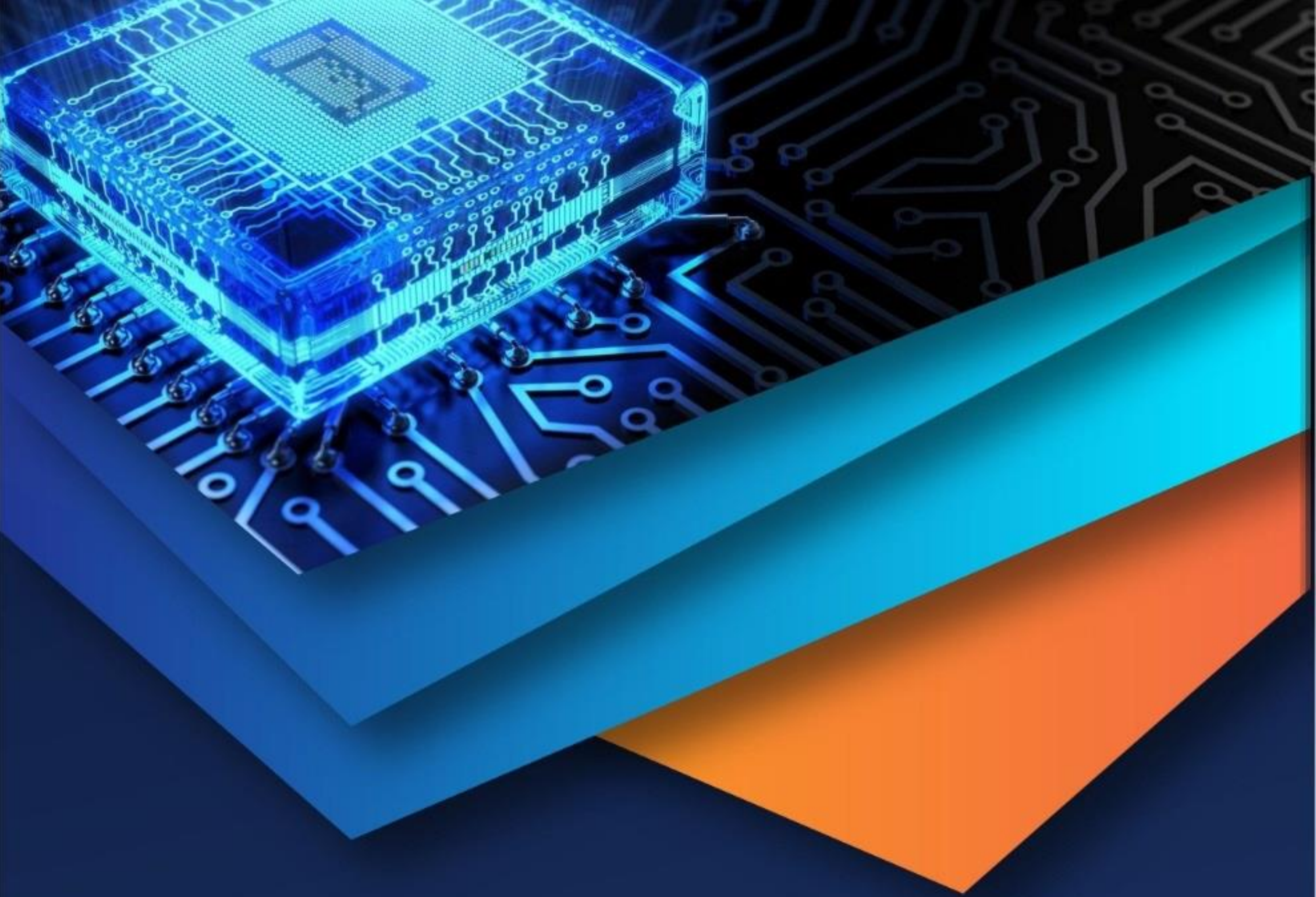
[6] S. Dhivyasri, et al., “An Efficient Approach for Interpretation of Indian Sign Language using Machine Learning,” in Proc. 2021 3rd Int. Conf. Signal Process. Commun. (ICPSC), 2021.

[7] L. K. S. Tolentino, et al., “Static sign language recognition using deep learning,” Int. J. Mach. Learn. Comput., vol. 9, no. 6, pp. 821–827, 2019.

[8] N. M. Adaloglou, et al., “A Comprehensive Study on Deep Learning-based Methods for Sign Language Recognition,” IEEE Trans. Multimedia, 2021.

[9] P. T. Krishnan and P. Balasubramanian, “Detection of alphabets for machine translation of sign language using deep neural net,” in Proc. 2019 Int. Conf. Data Sci. Commun. (IconDSC), 2019.

[10] T. Pariwat and P. Seresangtakul, “Multi-Stroke Thai Finger-Spelling Sign Language Recognition System with Deep Learning,” Symmetry, vol. 13, no. 2, p. 262, 2021.



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY