Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Real-Time American Sign Language Recognition on Edge Devices Using a Multilayer Perceptron

Muthu Lakshmi N.V.[1], Sunitha Kanipakam[2], Manjula K.[3], Prathyusha G.[4]

[1]Assistant Professor, [4]Academic Consultant, Department of Computer Science, SPMVV, Tirupati, India.
[2]Assistant Professor, Department of Law, SPMVV, Tirupati, India.
[3]Assistant Professor, Department of Applied Mathematics, SPMVV, Tirupati, India,

E-mail: [1]nvmuthulakshmi@spmvv.ac.in, [2]ksunitha@spmvv.ac.in, [3]manjula.karre77@gmail.com, [4]gp.spmvv@gmail.com

Abstract

The ultimate purpose of human existence is often considered to be finding happiness and satisfaction. People will live happily only when they communicate their feelings, emotions and opinions others. Furthermore, the, further right to freedom of expression is a fundamental right but people who suffer from hearing problems cannot communicate. This may lead to their isolation from societal activities. To address this issue, sign language came into existence, primarily to enable communication, education, employment and social inclusion. In this paper, an efficient methodology for American Sign Language (ASL) recognition using hand gestures is proposed. To find an efficient model, several experiments were conducted on ASL datasets which consisting of images, where each image represents signs for several words, alphabets, numbers, etc., Finally, the Multilayer Perceptron model with 2 hidden layers (1024,512) outperformed other models in terms of accuracy (99.3%) in static sign classification. These studies focused mainly on the recognition of legal words, as hearing people are unable to solve their legal issues because they cannot communicate effectively with their signs to the police or lawyers to raise their issues. In this methodology, the MediaPipe framework is used to extract landmarks from high quality sign-images, and pre-processing is done with normalization techniques, ensuring consistency in scale and orientation to generalize the model for various types of users in different environments. Experimental results demonstrated high classification accuracy, validated the model's performance and presented its potential for a range of practical applications in accessibility and human-computer interaction. A separate lightweight prototype was developed to evaluate optimal deployment models for real-time offline inference with quantized and non-quantized models on mobile hardware (CPU, GPU, and TPU) as additional work in this paper.

Keywords: American Sign Language, Hand Gestures, Media Pipe, Multilayer Perceptron, Mobile Hardware.

DOI: https://doi.org/10.36548/jiip.2026.1.007
Received: 03.12.2025, received in revised form: 03.01.2026, accepted: 17.01.2026, published: 27.01.2026

1. Introduction

Language and communication are the foundation for personal relationships within the family, friends and in professional settings. Clear communication helps in solving problems and managing conflicts, contributing to the overall happiness and emotional regulation of individuals. Thus, communication plays a vital role in leading a good life for any person. The United Nations recognized and established September 23rd as the International Day [1] of Sign Languages to highlight the crucial role of sign language in achieving human rights for Deaf people. Sign language is essential for human dignity and equal rights because it is a key to communication, enabling them to exercise their rights especially the right to education which is a fundamental right guaranteed under Article 21A and Article 14 provides the right to equality under the Constitution of India. Recognizing the need for and promoting sign language ensures the linguistic and cultural identity of the Deaf community which is a major fundamental aspect of human rights [2]and also promotes inclusive growth and a more equitable society. Access to sign language and education in sign language for deaf children at an early age is critical for their growth and development, enabling them to achieve educational and developmental goals. Moreover, it is crucial for human dignity.

Any person uses four types of communication modes: verbal, non-verbal, written, and visual. Among the four, verbal communication is the main source for conveying information, as it is very effective in communicating quickly and completely without any barriers. However, a person with a hearing problem may not be able to communicate with others because they cannot hear, and they may also be unable to talk due to various reasons, such as issues with their eyes, ears, or vocal cords resulting from health problems or accidents. Physically challenged individuals may sometimes be unable to communicate due to temporary or permanent vocal cord failure. In this context, sign language emerged as a non-verbal and visual mode of communication. It is a visual form of communication that incorporates facial expressions and body language, especially one or two-handed gestures.

The interpretation of sign language varies from country to country, as every country has several regions. In general, spoken language varies from region to region, reflecting local culture, history, and social factors. Sign language gestures are based on spoken language words, culture, lifestyle, and other factors. Therefore, gestures and body language can have different meanings in different regions.The main issue with sign language is the lack of universality. Approximately 300 sign languages were documented worldwide. The most well-known documented sign languages include American Sign Language, Indian Sign Language, French Sign Language, British Sign Language, and Chinese Sign Language.

The National Institute on Deafness and Other Communication Disorders (NIDCD), a part of the National Institutes of Health (NIH) focuses on ASL which is a standard and popular sign language, a complete natural language with its own grammatical structure, distinct from the English language. In the United States of America and many parts of Canada, American Sign Language (ASL) is the predominant sign language and is used by deaf and hard of hearing communities. More than 200 years ago ASL was formed from the intermixing of local sign languages and French sign language. Over time, ASL has differed from French Sign Language (LSF), but modern ASL and LSF have similar signs with equivalent meanings. In India, Indian Sign language shortly ISL is used but it is not unique to all regions within the country. ISL is a full-fledged language, capable of expressing complex concepts as effectively as spoken languages like English or Hindi. It possesses all the structural components that define a language and is not merely a system of gestures, but a distinct linguistic system with its own

grammar and vocabulary. The recognition of sign languages as complete languages emerged in the 1960s through research in the USA and the Netherlands, distinguishing them from animal communication systems.

In this paper, hand gesture recognition for sign language in ASL is considered.

Here, a comparison of prominent sign languages such as ASL, ISL, and BSL is presented.

1.1 Variation in Sign Language Gestures and Grammar of ASL, ISL and BSL

American Sign Language (ASL), British Sign Language (BSL), and Indian Sign Language (ISL) differ significantly in their gestures and structures. American Sign Language (ASL) requires only one-handed signs for alphabetical finger spelling; British Sign Language (BSL), on the other hand, uses a two-handed system. Indian Sign Language (ISL), however, uses ASL-like one-handed signs but varies in the shapes and movements. For the signs of numerical: in ASL, numbers from 1 to 5 are shown with the palm facing outwards; BSL uses two hands for counting numbers; ISL again follows ASL patterns but differs in higher numbers. Regarding grammar, ASL follows the topic-comment variation, which is different from English. ISL has a strong influence from its regional languages and differs across various states in India.

The differences are further highlighted by the gestural style. ASL heavily depends on facial expressions to emphasize and convey tone; BSL leverages body movements and two-handed gestures; and ISL is somewhat flexible and varies according to cultural practices. "Hello" in ASL is presented as a salute, whereas in BSL it is depicted as a wave, while in ISL, it is represented as a raised hand wave similar to our Indian traditional greeting. These distinctions illustrate that although sign languages serve the same purpose of visual communication, their gestures, grammar, and cultural adaptations make each unique.

Figure 1 shows how the sign for the word "JUSTICE" differs among ASL, ISL, and BSL.

Figure 1. Sign of Justice in ASL, ISL, BSL

The development of these automated sign language recognition systems greatly aids the deaf and hard-of-hearing communities in bridging their communication gaps and improving accessibility. Traditional Sign Language Recognition systems are often vision-based and heavily rely on complex and computationally intensive methods built on top of CNNs that internally perform image analysis, which involves training with pre-extracted high-dimensionality vectors.

Sign Language applications assist individuals with speech and language impairments, as well as those suffering from autism and other syndromes that hinder their ability to speak and communicate. Individuals facing temporary speech difficulties due to accidents or illness will also benefit greatly from these applications. Within a family that includes deaf members,

sign language further supports daily communication and strengthens interpersonal bonds. Sign language interpreters, educators, healthcare providers, the legal community, and others who work with the deaf and hard-of-hearing community need to be proficient in sign language.

Sign Language on Mobile Devices:Wearable sensors and computer vision based on cameras are the two prominent methods used for gesture recognition for sign language on mobile devices.

Video from the device's camera is analyzed by computer vision systems, such as those that use Google's MediaPipe framework, to identify landmarks and hand motions and translate them into signals. Sensors such as accelerometers for arm position and flex sensors for finger movement are used by wearable technology to record gestures. The goal of both approaches is to close communication gaps for those with speech or hearing impairments by processing the data using machine learning or artificial intelligence (AI) algorithms to convert sign language into text or voice. In this paper, user signs captured by computer vision on a mobile camera are considered.

This paper outlines a pipeline that employs the Google MediaPipe framework to extract hand landmarks, preprocesses these features to create a standardized input vector, and utilizes a Multilayer Perceptron (MLP) for robust classification. One of the primary objectives of this initiative is to develop a model that could be deployed on platforms with fewer resources or resource-constrained devices like mobile phones. Hence, TensorFlow Lite with a quantized variant is given priority over TensorFlow Keras H5 models.

The main contributions of this study are:

- A legal vocabulary in ASL dataset focused on
- Landmark-based representations and an explorative systematic evaluation of classical machine learning and MLP architectures, and
- Analyzing the performance of edge-device deployment and inference.

2. Related Work

Communication has been essential to the exchange of ideas and feelings throughout human history. Sign language serves as the main medium for specially-abled individuals [4] to connect with others through hand gestures. This paper focuses on classifying single- and double-handed Indian Sign Language using machine learning algorithms in MATLAB, achieving an accuracy rate between 92% and 100%.

We can broadly classify sign language recognition [5] models based on the type of input, the process of recognizing signs, and the technology used. Each is presented as follows:

Types of Inputs: There are three ways to receive input in sign language recognition models: vision-based, sensor-based, and hybrid models. Vision-based models capture signs through images or videos to detect hand and body gestures. The main advantage of this model is that it enables natural contact and is non-intrusive. Nevertheless, they encounter difficulties such as background fluctuations, occlusion, and sensitivity to lighting conditions. The second type of input uses wearable sensors, such as gloves and motion sensors, to recognize signs, and it works well even in poor lighting with high precision. However, it is expensive, and for long-

term use, it is not comfortable. The third model is hybrid, which combines vision-based and sensor-based methods to achieve better accuracy in recognizing signs. One advantage of hybrid models is their resilience to environmental changes. However, the computational complexity of this challenge is high.

Recognition Approaches: There are two recognition approaches: static gesture and dynamic gesture recognition. Static gesture recognition uses image classification models for gestures made in single alphabet signs, whereas dynamic gesture recognition involves motion over time, such as words or sentences. In this paper, computer vision-based inputs are utilized for research work. Many researchers have proposed various models using different machine learning techniques and have demonstrated their model performance through comparative analysis. The major challenge is finding complete and accurate datasets in any sign language. The dataset consists of two types of signs: finger spelling, which specifies alphabets and numbers, and signs for words or text.

Despite the existence of alternative sign languages [5], ASL is a thorough and organized visual language that conveys meaning through both manual and non-manual features. This paper focuses on sign language recognition in ASL and also discusses background noise, lighting variations, and real-time processing. Hearing-impaired people [6] select sign language as the best method for communication. There are various sign languages in the world based on their culture, language, tradition, and others which they follow. Moreover, sign language has different signs, and each sign is influenced by variations in hand form, motion profile, and the location of the hand, face, and body parts. Here, these movement profiles involve twisting the wrist, moving the hand in straight lines, such as upward, downward, or forward, circling the hand, bending or straightening the fingers, and also actions like bouncing the hand or reaching toward and touching a specific location. These movements are considered in forming, understanding, and distinguishing various signs. For example, in ASL, many noun-verb pairings are differentiated mostly by movement distinctions. Similarly, a noun is shown with a short or repeated movement, whereas a verb is shown with one long or continuous movement. The verb "sit" is indicated with a single continuous motion, but the noun "chair" is signified with a repetitive tapping action. In another instance, "airplane" and "fly" may have the same hand shape but different motions. By considering all the factors, a sign language recognition system can be designed effectively. Thus, recognizing sign language is a challenging task for people who need to communicate effectively and accurately.

In the United Kingdom, British Sign Language is used, so their sign language is based on British grammar, structure, and vocabulary but is different from spoken English. To convey meaning effectively, signs are also used based on body language, facial emotions, and hand shapes to convey meaning. In the year 2003, they officially recognized and made a legal sign language with the BSL Act 2022, which has had a significant impact on Deaf people's communication and culture. Many researchers are finding automatic interpretation of sign language, which can communicate effectively to other people using machine learning and deep learning models.

In this study [7], a continuous signed-letter identification system based on prior letter information and backhand-view rewound video sequences is presented. In comparison to conventional forehand and backhand view-based techniques, the method—which was implemented with a Leap Motion sensor and LSTM—achieved much higher recognition accuracy by resolving occlusion concerns and taking time-independent patterns into consideration. British Sign Language (BSL) is recognized through single and double-handed

motions in paper [8]. This study presents a CNN-based sign language recognition system. By extracting spatial and temporal characteristics from gesture photographs, the model improves the accuracy of identification, facilitating real-time translation and communication for the community of hearing-impaired people.

A multi-layer Convolutional LSTM is used by Mohammed, Adam A. Q., et al. in [9] for an automatic 3D skeleton-based dynamic hand gesture detection system. The framework accurately recognizes dynamic motions by capturing spatial and temporal information from 3D hand skeleton data. The method provides a reliable solution for real-time gesture detection in assistive technologies and human-computer interaction by combining convolutional layers with LSTM networks to efficiently model both temporal interdependence and spatial structure.

For applications such as sign language recognition and industrial automation, the authors in [10] proposed a hand pose estimation approach that precisely locates human hands and predicts their positions. Using this model, a Unity 3D game with six gestures was created. They demonstrated that the accuracy of the method was 95.63% using 21 hand joints. Animations, effects, and interactive Unity 3D elements were all created using Blender software.

The FFV-Bi-LSTM algorithm [11], which combines Fast Fisher Vector and Bi-directional LSTM for dynamic sign word recognition, is proposed in this study. It converts frames into high-dimensional vectors by utilizing the Leap Motion Controller's 3D hand skeleton and orientation characteristics. The technique enhances performance on the LMDHG and SHREC datasets by up to 3.19% and reaches 98.6% accuracy on ASL.

In paper [12], cutting-edge deep neural networks are used, and the research provides a thorough comparative analysis of computer vision-based sign language recognition. It discusses different pre-training strategies, evaluates performance on several public datasets, develops the first RGB+D Greek sign language dataset with both sentence-level and gloss-level annotations for video sequences, and presents novel sequence training criteria inspired by speech and text recognition.

This research suggests a deep learning-based method [13] that uses LSTM and GRU networks to recognize words from gestures. The model, which consists of a single LSTM layer followed by GRU, has been tested on the IISL2020 dataset containing isolated Indian Sign Language video frames. It achieves approximately 97% accuracy across 11 signs, providing a useful way to assist persons who are speech- or hearing-impaired in communicating.

Duraimutharasan and Sangeetha [14] have utilized machine learning approaches such as CNN and Computer Vision (CV) for identifying Indian Sign Language words. Their work depicts the value of using a supervised machine learning approach to understand the complex patterns in ISL and focuses on improving accessibility by locating the decision limits in N-dimensional space for gesture classification to help the hearing impaired.

Another research work [15] proposed a unique technique that uses a webcam to capture photos and shows the meaning of the respective sign language signs as text on the associated display device. This research leverages images of hand gestures and a Convolutional Neural Networks (CNNs) approach to achieve this functionality. After training the model, the system identifies input gesture images by matching parameters and returns the interpreted language signs as meaningful text.

Kumar, Ashutosh et al. [16] have presented a real-time sign language recognition system that works for American Sign Language (ASL) using MediaPipe and CNN. MediaPipe Solutions provides a suite of libraries and tools for quickly applying artificial intelligence (AI) and machine learning (ML) techniques in our web or mobile applications. MediaPipe efficiently identifies hand landmarks, while CNNs classify the captured data into gestures. ASL users and non-signers can communicate easily thanks to this integration, which improves accuracy and lowers computing complexity while providing a low-latency, effective assistive technology for real-time gesture-to-text conversion.

A machine learning technique introduced by Srivastava et al. [17] was developed to interpret hand gestures and translate them into comprehensible text. The research showed that gesture classification was performed with a remarkable degree of accuracy and that machine learning could serve as a facilitator of communication and an accessibility tool for people with speech or hearing impairments by bridging the gap between them and others.

To recognize alphabets in sign language, a deep learning-based method is presented by Josef, A. and Kusuma, G. P. [18], using Bayesian optimization to adjust model parameters. Hand gesture identification accuracy is significantly improved in this model, demonstrating how deep learning and optimization techniques can be combined to provide reliable and efficient sign language recognition. A framework for sign language recognition that combines voice synthesis with Long Short-Term Memory (LSTM) networks was presented by Kumar, Mukesh, et al. [19] to facilitate communication for those who are hard of hearing. The system uses deep learning-based assistive technology to bridge the communication gap by recording sign motions, translating them into intelligible speech, and using LSTM for sequential feature learning. This allows for real-time translation.

In paper [20], a real-time caption generation framework for ASL that combines YOLO and LSTM was presented by Jana, Urmi, et al. The YOLO algorithm is used for accurate and fast hand gesture detection, while LSTM processes temporal sequences to interpret signs and turn them into meaningful text captions. This integrated approach improved accuracy and efficiency, facilitating smooth communication between ASL users and non-signers.

Authors Fertl, Castillo, Stettinger, Cuéllar, and Morales presented in paper [21] an extensive survey on device-free hand gesture recognition (HGR) methods designed for edge devices. The authors highlight numerous sensor modalities, such as vision, radar, WiFi, mobile networks, and ultrasound, in addition to pre-processing techniques such as stereo vision and spectrograms. They provided a classification of approaches from decision trees to transformers, while also acknowledging the hurdles caused by algorithmic complexity, hardware limitations, power consumption, latency, and memory usage. The authors pointed out the issues, the associated trade-offs, and the strategies for achieving fast and efficient HGR in devices under power, memory, and processing capacity constraints.

In paper [22], the authors introduced a real-time system that can recognize two-handed Indian Sign Language (ISL) gestures using MediaPipe and machine learning techniques. The proposed method is based on a two-hand recognition scenario where MediaPipe Hands are used to accurately extract the landmarks by recognizing both the left and right gestures simultaneously. A Convolutional Neural Network (CNN) subsequently takes care of processing and classifying the landmarks so that the system can master a set of static ISL signs. The pyttsx3 text-to-speech engine is then deployed to convert the recognized signs into spoken words, thus improving accessibility and facilitating communication between the public and the hearing-

impaired. The system ensures a scalable and efficient approach for real-time ISL recognition with high accuracy in different plant operation situations and environments.

A sign language recognition system is proposed by Donepudi, S., Divya, et al. [23] to improve communication for people who have speech problems. The system records and interprets hand gestures using computer vision and machine learning. It promotes inclusive communication with the general public by acting as an assistive tool that converts these gestures into meaningful communications.

3. Methodology

In this paper, a machine learning model isdeveloped to assist hearing-impaired persons in communicating with legal professionals to make any complaint at the police station or to file a case against a person who is committing injustice in any way against them. As they cannot communicate directly, they can convey their messages through sign language using their hands. These hand gestures are captured and fed into the model as input, which translates them into words, allowing the receiver to understand what the hearing-impaired person wants to communicate.

A methodology is proposed to develop a real-time, vision-based sign language recognition system using hand gestures for edge devices. In this model, the MediaPipe framework extracts key hand landmarks from images. The identified features are then normalized to maintain consistency in scale and orientation and to generalize the model across different users and environments. Finally, the processed landmark data is used as input to an efficient Multilayer Perceptron (MLP) classifier. The following section explains each step in detail.

3.1 Selection of Quality Images for sign Languages

As there are several sign languages, American Sign Language (ASL) is a fully developed, natural language with its own grammar, distinct from English. It is conveyed through hand movements and facial expressions. ASL is the primary language for many Deaf and hard-of-hearing individuals in North America and is also used by some hearing people.

Although they are all natural visual languages used by Deaf people, American Sign Language (ASL), British Sign Language (BSL), and Irish Sign Language (ISL) are grammatically different from one another. ASL is not mutually intelligible with BSL or ISL, even though English is spoken in the United States, the United Kingdom, and Ireland. While BSL and ISL have their own distinct origins and development, ASL is actually more closely related to French Sign Language (LSF) because of historical influences. Each sign language has a distinct structure and usage due to these variations in vocabulary, grammar, and syntax.

ISL is used in both the Republic of Ireland and Northern Ireland, BSL is used across the United Kingdom, and ASL is the predominant sign language in the United States and some parts of Canada. It's interesting to note that ISL has some characteristics in common with both French Sign Language and BSL, which results in a unique linguistic combination. Furthermore, the manual alphabets are different: BSL and ISL employ two-handed alphabets, while ASL utilizes a one-handed alphabet. These sign languages are not interchangeable even if they have comparable communicative functions and represent the historical and cultural backgrounds of their respective locations.

In this paper, ASL is considered for research. The sign language has a set of images to describe letters, numbers and words. However, not all images are taken for research as some images may have impurities; therefore, those images must be eliminated and this process is done manually in this work.

In order to build a model that can assist in courtroom discourse, we had to create a robust preprocessing pipeline that goes beyond conventional data cleaning processes. Our rigorous processing pipeline started with a vocabulary of 150+ legal / courtroom related and commonly used terms which accounted for ~6000 initial images curated from the ASL image corpus. Each sign was manually cross-verified using ASL dictionaries and instructional resources such as YouTube tutorials. This verification and validation process allowed us to eliminate ambiguities and build a robust dataset.

After the initial steps, for each word in the image dataset, thirty image samples were collected low quality or blurry images, as well as images with improper angles or incorrect sign language postures, were discarded. The samples that were pristine and demonstrated visual clarity were retained. This resulted in a tiny, substantially reduced yet highly reliable dataset of 1500 images. This forms our foundation for model training and testing ensuring that classification is grounded in terms of correctness and quality. Figure number 2 - demonstrates the process of preparing the pristine dataset.

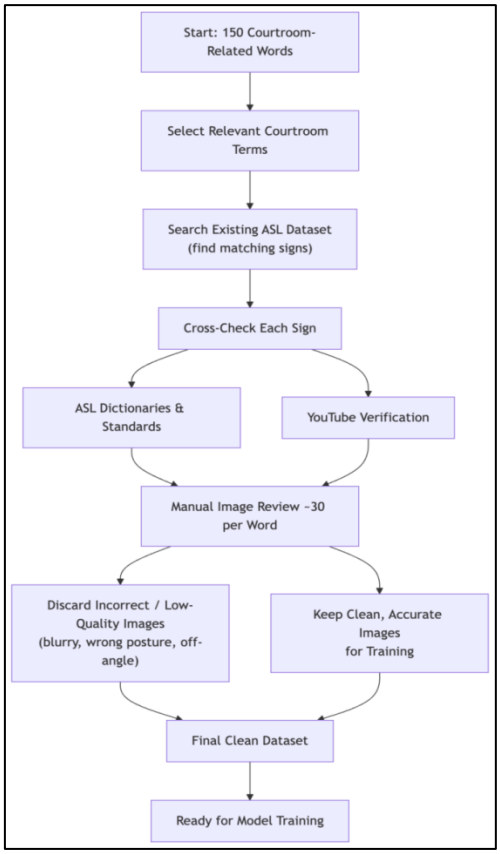

Figure 2. Selection of Quality Images for ASL Signs related to Legal Words

3.1.1 Dataset Construction and Splitting

The dataset was collected in a multi-step manner using a refinement process. First, around 6000 ASL images from the internet were gathered, and then 1500 images were retained

through manual verification and quality filtering, corresponding to 150 legal signs. Each class has roughly 10 representative examples. The final dataset was divided into two parts, training and testing, using an 85:15 ratio and a fixed random seed of 42 to ensure that the process can be reproduced. All reported results in the experiments are based on this finalized dataset unless mentioned otherwise.

3.2 Feature Extraction

The initial and most critical stage of our pipeline is the extraction of key features from the visual data. Rather than processing raw image pixels, which requires a large number of parameters and is sensitive to lighting and background noise, we employ the MediaPipe Hands solution. MediaPipe is a cross-platform, open-source framework for building multimodal machine learning pipelines. Its Hands module is specifically designed for high-fidelity hand and finger tracking, capable of inferring 21 distinct 3D keypoints (landmarks) for each hand from a single image.

There are three major steps in finding key points from the image, which are as follows:

Step 1:

First, the raw images are converted from BGR (Blue-Green-Red OpenCV) color space to RGB (Red-Green-Blue MediaPipe) color space. This is essential because MediaPipe's models are predominantly trained in RGB color space, hence this translation helps to retain the integrity of visual features and enables accurate detection of hands and their respective coordinates. Figure 3 shows raw image for the word "camera" and the changed image colors.

Figure 3. Raw Image for the word "Camera" and Conversion of Colors

Step 2: Palm Detection

The Blaze Palm model, a lightweight single-shot detector, identifies the location of the hand first within the entire input image. This model is mobile optimized for its performance and is robust against variations in hand size and position.

Step 3: Hand Landmark Localization

After the palm has been detected by the framework, the image is cropped to show only the hand region. Then, a second neural network, which is more accurate, is used to perform regression on the cropped area in order to determine the exact 3D coordinates (x, y, z) of the 21 landmarks. These landmarks are situated on the palm, thumb, and the other four fingers (index, middle, ring, and little).

3.2.1 Preprocessing Steps

Let us assume that I is the image that the Mediapipe hands model is processing. The Hands Model returns 21 hand landmarks per hand, totalling 42 landmarks in total for 2 hands. 3 co-ordinates per landmark equal a 126 value vector that will be returned. These preprocessing steps are depicted in the flowchart in Figure 4.

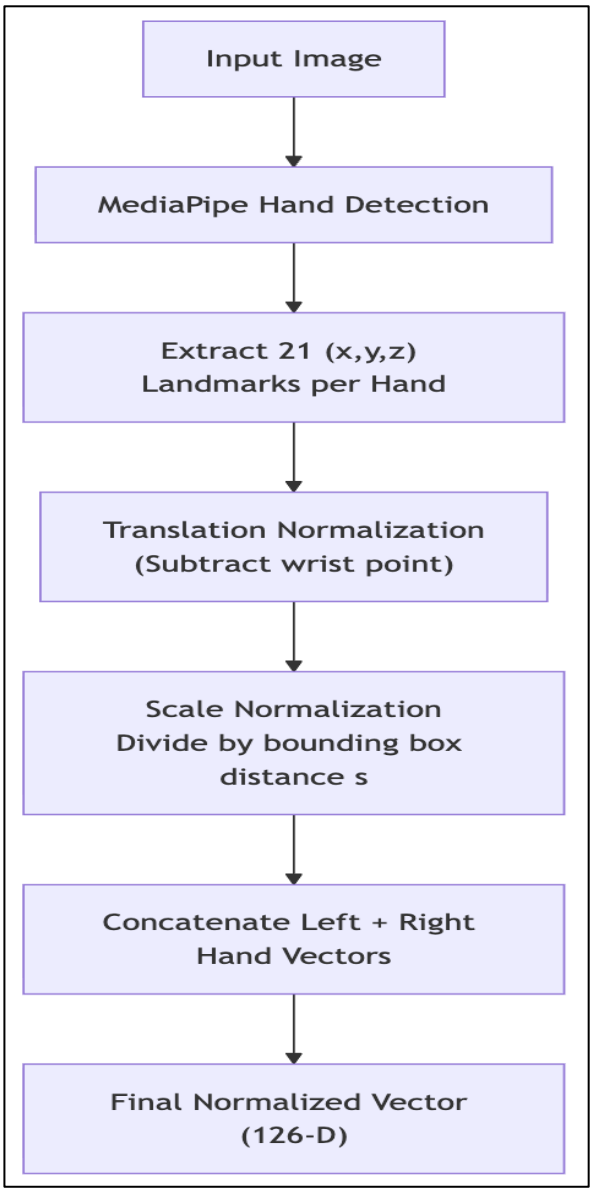

Figure 4. Flowchart for Pre-Processing Steps to find Key Points

For one hand, Number of landmarks N = 21 be:

$$L = \{(x_i, y_i, z_i) \mid i = 1,2, \dots, N\}$$

If two hands are detected by the model, then the combined feature vector would be:

$$v = [x_1, y_1, z_1,\ \ x_2, y_2, z_2, \dots, x_{2N},\ \ y_{2N}, z_{2N}]$$

If only one hand is detected by the MediaPipeHands model, then the second hand will be padded with zeros in the vector for all 21 landmarks – [(0,0,0),(0,0,0),….]:

$v = [v_1, 0_{3N}]$.

The landmarks are normalized w.r.t first landmark (x1,y1,z1):

1. Translation normalization (centering):

$$\hat{L}_i = L_i - L_1 = (x_i - x_1,\ y_i - y_1,\ z_i - z_1)$$

2. Scale normalization (to achieve size invariance):

Say,

$$v_{max} = (max_{\ i}\ x_i, max_{\ i}\ y_i, \underset{i}{max}\ z_i)$$

$$v_{min} = (min_{\ i}\ x_i, min_{\ i}\ y_i, min_{\ i}\ z_i)$$

Then the scaling factor becomes:

$$s = \| v_{max} - v_{min} \|_2$$

Now the final normalized coordinates are as follows:

$$\tilde{L}_i = \frac{\hat{L}_i}{s}$$

3.2.2 Final Feature Representation

The normalized landmarks for both the hands are concatenated into one final feature vector, which is

$$x = [\tilde{x}_1, \tilde{y}_1, \tilde{z}_1, \tilde{x}_2, \tilde{y}_2, \tilde{z}_2, \dots\dots\dots, \tilde{x}_{2N}, \tilde{y}_{2N}, \tilde{z}_{2N}]$$

where

$$x \in \mathbb{R}^{6N} (for\ two\ hands, 126\ features\ total)$$

In Figure 5, the hand gesture has the key points indicated.

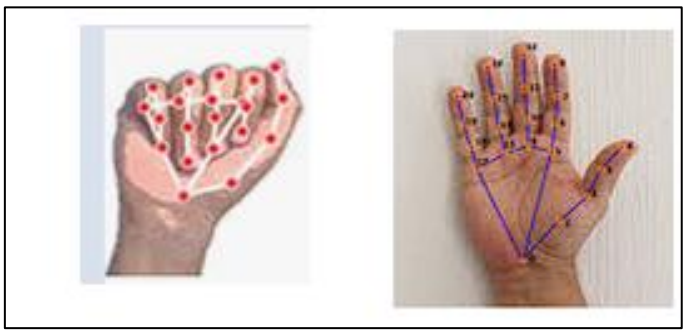

Figure 5. Landmarks, Key Points from Hand Gesture

In Figure 6, the sign for the word 'camera' in ASL is shown along with the corresponding landmarks of the detected palms.

Figure 6. Land Marks Detected from Palms

This three-step technique is computationally effective, since the intensive palm detection task is carried out only on the entire image, while the landmark detection is conducted on a small, cropped area. The result is a set of landmark coordinates for every hand that has been spotted.

3.3 Data Preprocessing

The unprocessed landmark coordinates that MediaPipe provides are related to the camera's field of view and the size of the detected hand. These raw values are not the most suitable for a machine learning algorithm because a close-up gesture would be much larger and have different absolute coordinate values than a similar gesture that is farther away. Thus, a very important normalization step is taken to obtain invariance regarding scale, position, and orientation.

The MediaPipe algorithm provides the landmark coordinates relative to the camera's field of view and the size of the hand gesture being detected. We also understand that the same gesture might yield different absolute values depending on the distance of the hand from the camera. Hence, these coordinates cannot be fed into machine learning models directly. Thus, we have to employ normalization techniques to solve the problems of scaling, positioning, and achieving orientation invariance.

The normalization process consists of two primary transformations:

- **Translation to Origin:** The coordinates of all 21 landmarks are translated so that the wrist landmark (point0) is at the origin (0, 0, 0). This effectively canters the hand pose and makes the feature vector independent of the hand's absolute position in the frame.

- **Scale Normalization:** The translated coordinates are scaled using a factor based on the overall size of the hand. Specifically, this scaling factor is computed as the Euclidean distance between the minimum and maximum coordinate values along each axis, which defines the hand’s bounding box. By dividing all coordinates by this factor, gestures of varying sizes are normalized to a consistent representation.

This normalization yields a standardized, 126-dimensional feature vector (2 hands x 21 landmarks x 3 coordinates/landmark). If only one hand is detected, the landmark data for the second hand is represented by a zero-filled vector to maintain a consistent input size for the model. This vector is then paired with its corresponding sign language label. The string labels are first converted into integer representations using the Label Encoding technique, and then one-hot encoding is applied for categorical data.

Zero-padding is applied exclusively as a representation of the absence of a second hand, rather than as an artificial movement, which is required by feedforward classifiers to maintain a fixed-dimensional input representation. Even though it is a good approach for reducing the problem of shortcut learning, it will still be replaced in future work with explicit hand-presence masking.

3.4 Methodology for Sign Language Classification

After feature extraction and data preprocessing, cleaned data will be ready to be fed to the machine learning model. A Multilayer Perceptron (MLP), a type of feedforward artificial neural network tailored for classification, forms the foundation of the gesture recognition system.

Although advanced architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are widely used in computer vision and time-series tasks, the MLP is considered the most appropriate classifier for reliable sign language classification. The selection of the Multilayer Perceptron (MLP) classifier over Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) models for sign language recognition is justified as follows.

MLP vs. CNN: CNNs are primarily designed to process raw pixel data by learning spatial hierarchies of features through convolutional and pooling layers. In this pipeline, =feature extraction and engineering are already handled by MediaPipe, which provides a clean, structured landmark vector. Feeding this pre-processed, non-grid-like data into a CNN would be inefficient and would not leverage the CNN's core strength of learning spatial relationships from raw imagery. The MLP, on the other hand, is perfectly suited for classifying structured, tabular data like our normalized landmark vectors.

MLP vs. RNN (LSTM): Recurrent Neural Networks, with LSTMs in particular, are intended for sequential data processing while preserving the temporal dependencies that exist among them, thus making them the best contenders for dynamic gesture recognition (e.g., words that incorporate motion). However, the current study is limited to static hand gestures (e.g., fingerspelling the alphabet). There is no necessity for temporal modelling because each picture and its landmark vector can be regarded as isolated instances. The forward nature of the MLP is adequate and less computationally expensive c for this static classification problem.

The selection of a basic multi-layered MLP for this structured data problem illustrates an essential principle in machine learning: picking the proper tool for the job. With MediaPipe, feature engineering is offloaded to a great extent. The neural network task is thus simplified, allowing even a simple MLP to achieve brilliant performance with a smaller model size and quicker inference times. The resulting model is, therefore, very suitable for deployment on edge devices.

The MLP model is constructed through the Keras Sequential API, which provides an easy way of layering the units. Its structure includes an input layer, two hidden layers, and an output layer at the end.

- **Input Layer:** The input layer takes a normalized vector comprising 126 features. The size of this input is defined by the number of features, which, in this model, is 126.

- **Hidden Layers:** There are two dense layers with 1024 and 512 neurons, respectively, and both are fully connected. The ReLU activation function is applied in both hidden layers. ReLU is chosen for its computational efficiency and its ability to mitigate the vanishing gradient problem, which can occur in deep networks.
- **Output Layer:** The final dense layer has a number of neurons equal to the total number of sign classes. The softmax activation function is applied to this layer, which converts the raw output scores into a probability distribution over the classes. The class with the highest probability is selected as the model's final prediction.

Then, the model is compiled with the Adam optimizer, which is a popular and effective optimization algorithm that adapts the learning rate during training. The Cross-Entropy Loss function is used to compute the loss, which is the standard choice for multi-class classification problems where the target labels are one-hot encoded. The Multi-Layer Perceptron architecture can be described in Figure 7(a) and 7(b). These two figures illustrate the major steps in the sign language gesture recognition system for ASL data.

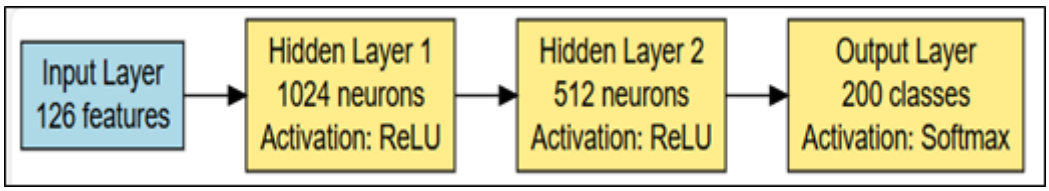

Figure 7(a). Sequence Process of Multi-Layer Perceptron Architecture for Sign to Text

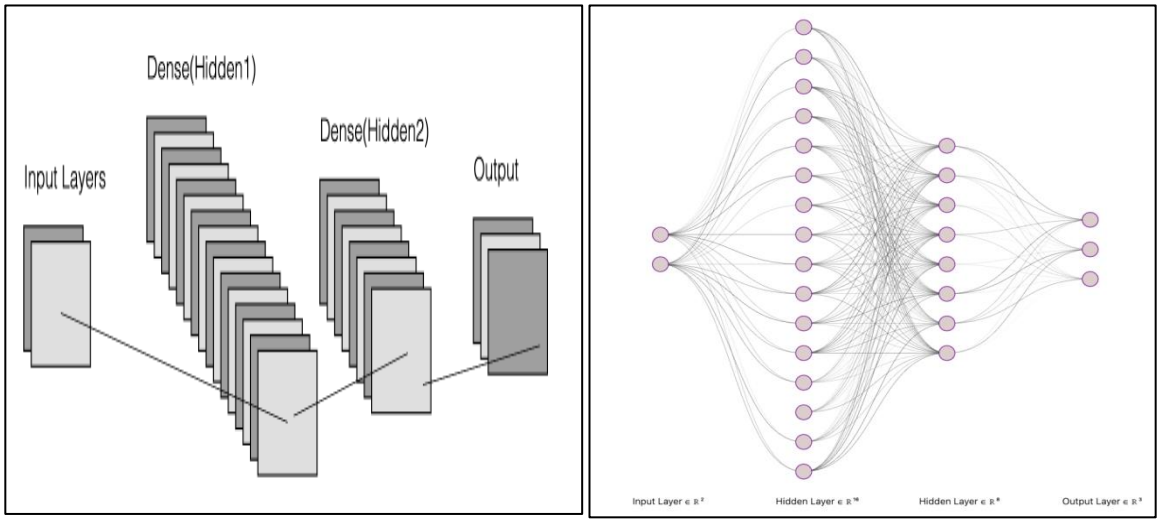

Figure 7(b). AI Model-Multi-Layer Perceptron Architecture for Sign to Text

Multi-Layer Perceptron (MLP) with hidden layers of size (1024, 512) is used as the final model for American Sign Language (ASL) recognition because it consistently outperformed other candidates in terms of accuracy and generalization. The performance of the proposed model is determined by evaluating key metrics such as accuracy, precision, recall, F1-score, etc.

3.3.1 Clarification on Static Image Modelling and Mobile Prototype

The research presents a model based on the multi-layer perceptron (MLP) algorithm that is only trained and assessed using static images derived from MediaPipe hand landmarks.

The live camera input used for the mobile demonstrations is processed in a frame-by-frame approach without considering the time factor or following any sequence modelling. Furthermore, a distinct video-based prototype was developed with Google Teachable Machine just to show the real-time potential on mobile devices. This prototype is separate from the suggested MLP model and is not part of the experimental evaluation or comparative results discussed in this paper.

4. Experimental Results and Analysis

The MLP model is trained on the prepared dataset using a standard supervised learning approach. The data is split into a training set and a testing set to evaluate the model's generalization capabilities on unseen data. The model is trained for 25 epochs, and its performance is validated on the test set after each epoch.

After training, the model is saved in the native Keras format. For real-world deployment on mobile or embedded devices, the model is converted to the TensorFlow Lite (TFLite) format. The TFLite converter optimizes the model by applying techniques like quantization, which reduces the model size and improves inference speed without significant loss of accuracy. The final TFLite model, along with the saved categorical encoding file, can then be integrated into an application for on-device, real-time sign language recognition.

4.1 Classifier Families and Performance Analysis

In order to measure the recognition of American Sign Language (ASL) hand signs, a vast number of different classical machine learning classifiers and neural architectures were used in the experiments.

The models included decision-based ensembles, distance-based methods, kernel machines, probabilistic classifiers, gradient-boosted trees, and multi-layer perceptrons(MLPs).

The goal of this study was to not only highlight the best-performing models through comparison but also to see how the different algorithmic households make use of the visual landmark features.

- **Logistic Regression (LR):** This is a linear classifier with L2 regularization and represents a baseline. The moderate performance (~73%) indicates that it is not very easy to separate ASL landmark data with a linear decision boundary, and thus, a more expressive framework is required**.**
- **Decision Trees and Ensembles (Random Forest, Extra Trees):** Decision Trees capture the non-linear interactions of features by recursively partitioning the feature space. The evidence shows that random forests and extra trees, both being ensemble variants, combine multiple trees to decrease the variance. These models achieve nearly perfect accuracy (close to 1.0), which reflects that the combination of landmark-based features through ensemble averaging is very discerning and may lead to overfitting as well**.**
- **K-Nearest Neighbors (KNN):** KNN assigns a class by comparing the distance of the feature point to other points. Results indicate that it performs very well (~92-98%) with small values of k, meaning the samples of the same class are distributed

quite closely in the landmark space. Thus, Weighted KNN obtains the maximum scores which indicates that the weighting of the local neighborhood of the case has been effective in resolving ambiguity.

- **Support Vector Machines (SVM):** SVMs using linear kernels exhibit impressive performance (45%) but strong at C=10 (~90%). This inconsistency highlights the need for proper kernel parameter tuning. Polynomial kernels do not perform as well, which might be due to the model fitting high-degree interactions in limited training data.
- **Naïve Bayes:** The Gaussian Naïve Bayes method considers features to be conditionally independent. Even with this unrealistic assumption, it manages to achieve ~97% accuracy, indicating that class distributions are still quite separate under Gaussian assumptions.
- **XGBoost:** XGBoost, which is a gradient-boosted tree framework, achieves 100% accuracy and thus proves its capability in capturing complex non-linear patterns with the help of boosting.

4.2 Multi Layer Perceptron Architecture

Multi-Layer Perceptrons (MLPs): Different MLP architectures aim to show the impact of model depth, width, and activation functions on classification through neural networks.

- Shallow MLPs (single hidden layer of 64–256 neurons) improve steadily with increased capacity, rising from ~72% to ~91%.
- Medium-depth MLPs (two hidden layers, e.g., 256–128) reach ~97%, confirming that modest depth is sufficient for this dataset.
- Deeper MLPs (e.g., 512–256–128) further approach 99%, rivaling ensemble methods.
- Activation functions significantly affect performance: ReLU consistently outperforms tanh and logistic, the latter collapsing to near-random performance in deeper settings due to vanishing gradients. Very wide MLPs (1024–512) achieve the highest neural accuracy (~99.3%), suggesting that over parameterization combined with sufficient regularization allows excellent generalization.

These results reinforce well-established findings in deep learning literature: while logistic and tanh are biologically inspired, ReLU has become the de facto standard for modern neural architectures due to its stability and scalability.

Compared to other models, MLP models have advantages in learning non-linear feature interactions well, which random forests cannot capture. Generalization is better when trained with augmentation of different hands and rotations; lighting also offers a differentiable pipeline that integrates well with deep-learning extensions.

Here, the graphs in Figure 8, Figure 9 and 10 show a comparison of the top 10 machine learning models.

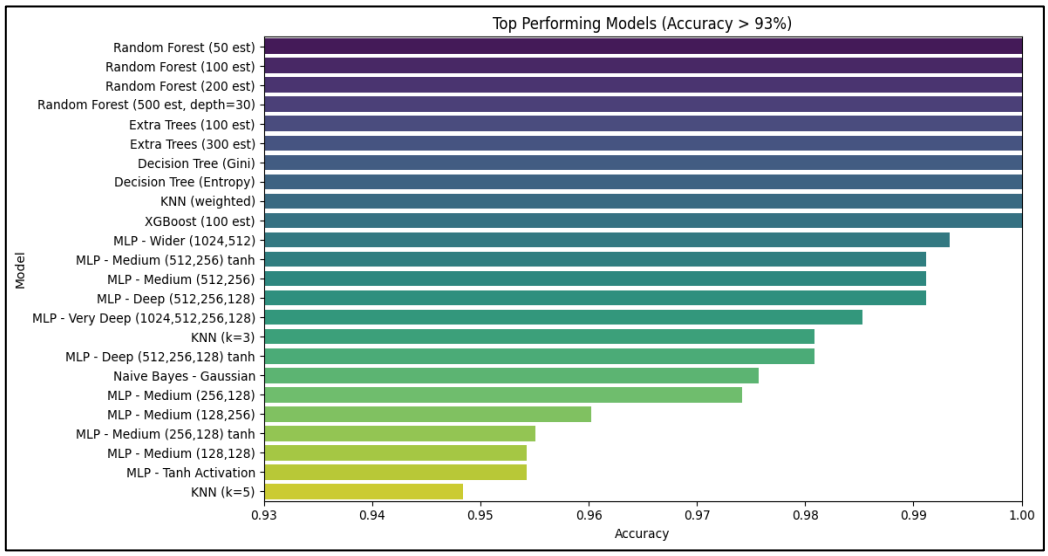

Figure 8. Various Models and Its Accuracy >93%

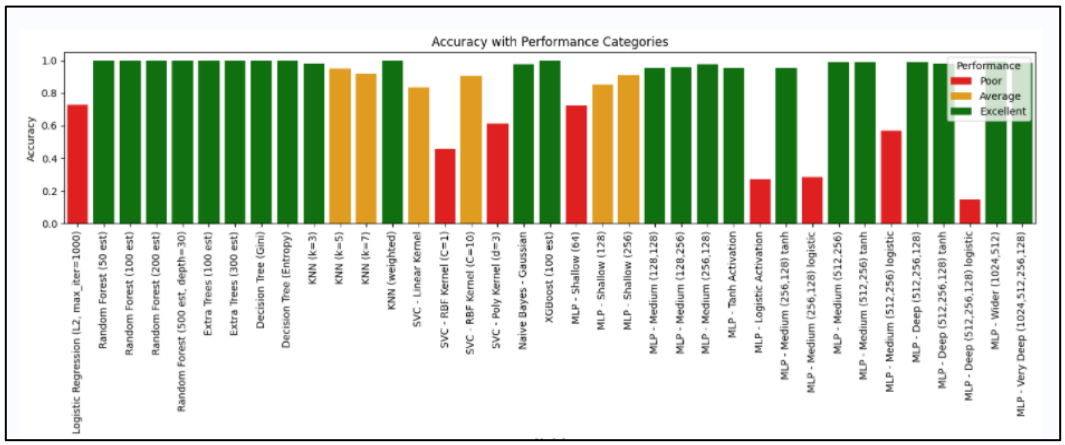

Figure 9. Comparison of All Models by their Accuracy

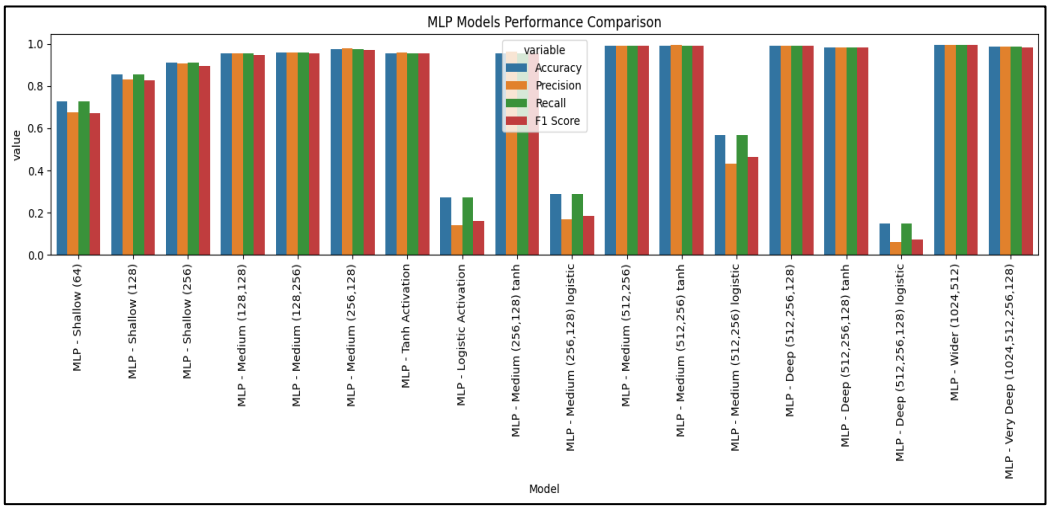

Figure 10. MLP Models for Hand Gesture Recognition with Different Measures

Random Forests achieved strong results but rely on partitioning feature space into discrete rules. While effective for small datasets, they struggle to extrapolate to unseen hand variations for new users, lighting, and variations in orientation. KNN and Decision Trees are simple and interpretable but scale poorly with large datasets and often lead to overfitting. SVMs

performed moderately but required heavy kernel tuning and had slower training times. Multi-layer perceptrons were able to capture the non-linear features that other models could not capture and have proven to generalize better. Although some classifiers reached 100% accuracy, in a curated dataset, such results could be a sign of memorization rather than generalization. Hence, the models were not only judged by their accuracy, preference was given to those architectures that showed consistent performance, smooth decision boundaries, and suitability for use by new users and on edge devices.

5. Experimenting with Real-Time Sign Language Classification for Mobile Deployments

After the success with the static image-based ASL processing model using machine learning and deep learning techniques in the initial phase under controlled conditions, the next step is to bundle the model for deployment onto mobile and edge devices for offline inference. Hosting the model on the cloud for inference leads to latencies due to limited network bandwidth and restricts usage to locations with reliable internet access. To facilitate deployment on mobile devices, TensorFlow Keras H5 and TensorFlow Lite (TFLite) conversions were explored, where quantized (8-bit INT), and floating point (32 bit Float) model variants were available. Experiments were carried out to identify the best variant for offline inference in terms of inference speed, accuracy retention and model size. The models performance has also been calibrated across multiple hardware backends including CPU, GPU, and TPU.

The highest precision TensorFlow (Keras H5) models and TFLite floating point (FP32) models have consistently delivered highest accuracies but with the trade-off of the largest model sizes and slow inference times. The 8-bit quantized TFLite model which is 67% smaller than the Keras model, provides the fastest inference with single digit millisecond inference times (<8ms) across CPU, GPU and TPU hardware which is over 95% faster compared to Keras H5 model variants.

In order to assess the different mobile deployment features like quantization, model size, and inference latency a separate lightweight prototype using Google Teachable Machine was built. This prototype has been trained only on a small number of hand gestures and color classes and was used only for the experimentation with quantized and non-quantized models on mobile hardware. The prototype does not rely on the proposed MLP-based classifier and does not affect the experimental results or the performance comparisons presented in this paper.

Successful prototyping of frame-wise classification for live-video feed with a limited set of static hand signs and colour classification has been carried out using Android Studio and thequantized variant of the TensorFlow Lite family, which occupies minimal storage footprint and delivers the fastest inference without significant compromise in accuracy for mobile deployment without reliance on the cloud.

This successful exploration of integrating a classification model with an Android device's live feed and its results lays a strong foundation for our future work to support real-time translation across multiple sign languages, including Indian Sign Language (ISL), American Sign Language (ASL), and British Sign Language (BSL). This progression aims to establish a fully integrated classification system optimized for mobile platforms.

The below Figure 11 provides screenshots of the signs "Paper" and "Rock" in the mobile app.

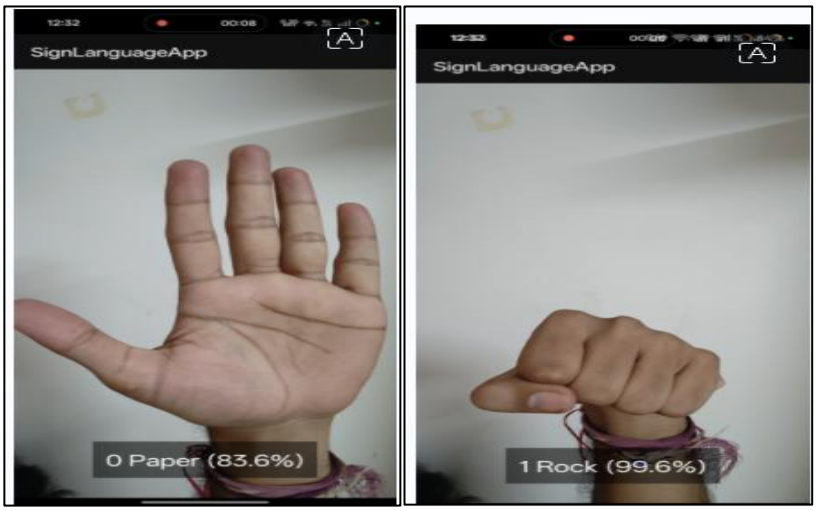

Figure 11. Sign to Text for the Word Paper and Rock

Model sizes in terms of MB of these experiments are shown in Figure 12.

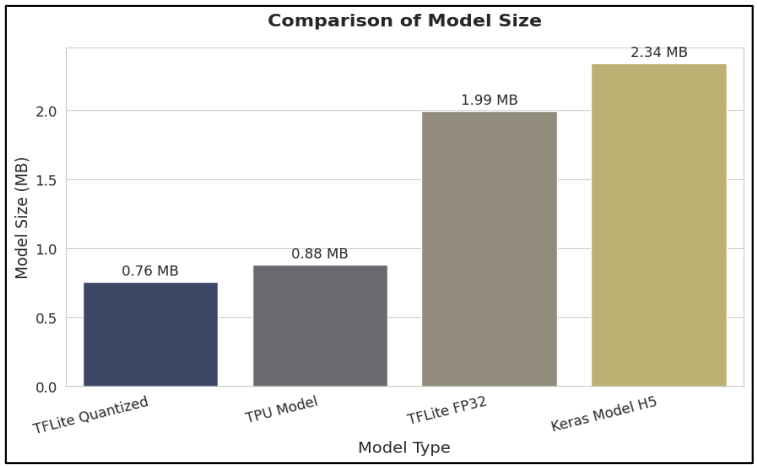

Figure 12. Keras H5, TFLite FP32 and TFLite Quantized Models and Its Size

Average inference time observed in various experiments on three hardware devices is shown in the following Table 1.

Table 1. Comparison of Average Inference Time Over CPU, GPU, TPU

S.No.	Name of the Model	Average Inference Time (in Milliseconds)	Device
1	Keras H5	233.409214	CPU
2	TFLite FP32	3.461822	CPU
3	TFLite Quantized	7.216819	CPU
1	Keras H5	381.009849	GPU
2	TFLite FP32	2.469397	GPU
3	TFLite Quantized	5.323235	GPU
1	Keras H5	113.787206	TPU

2	TFLite FP32	1.920859	TPU
3	TFLite Quantized	3.234116	TPU

The graphs corresponding to the above inference data are depicted in Figures 13(a) and 13(b) are as follows.

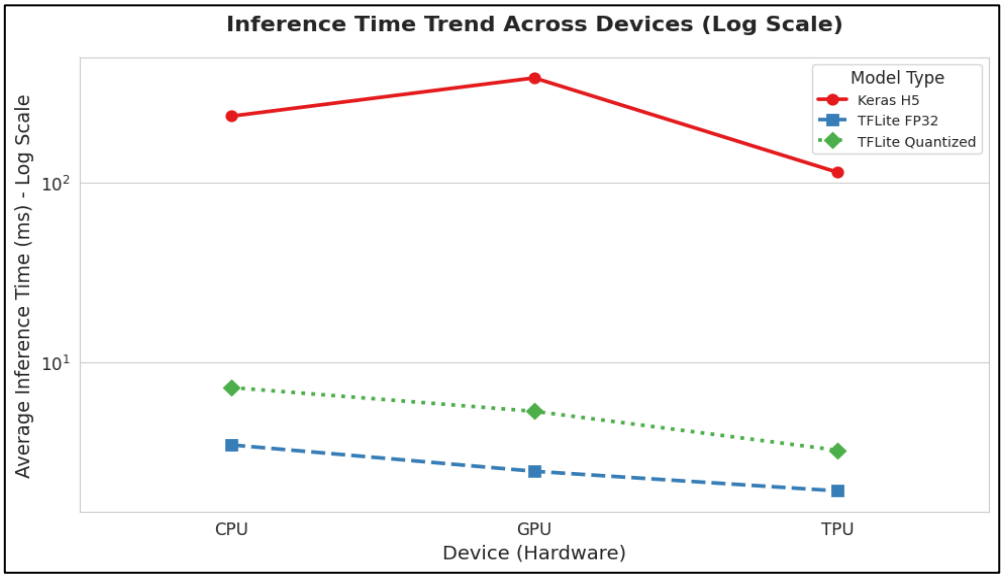

Figure 13(a). Line Graph-Average Inference Time of Keras H5, TFLite FP32, TFLite Quantized over CPU, GPU, TPU

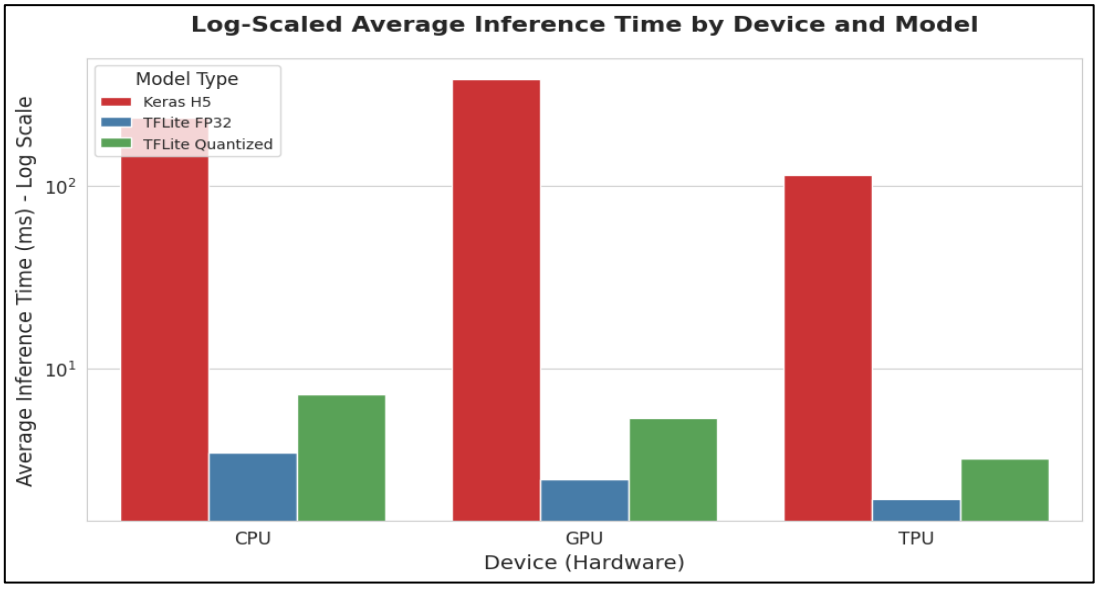

Figure 13(b). Histogram-Average Inference Time of Three Models over CPU, GPU & TPU

6. Conclusion

Sign Language Recognition (SLR) systems play a crucial role in enhancing communication for individuals with hearing or speech impairments, promoting independence, security, and social inclusion. They act as an essential bridge between hearing and non-hearing communities, enabling individuals with hearing impairments to independently register

complaints and communicate in legal or public contexts. This paper focuses on recognizing legal vocabulary in American Sign Language (ASL) to assist in preparing documents and facilitating interactions with legal professionals. Various traditional classification models and neural network architectures were designed and evaluated using training, validation and testing phases with different parameters, where MLP based models tended to generalize better. The models made use of images featuring ASL signs and mainly focused on the correct identification of legal terms that ranked among the top 150. These terms were gathered from various web sources including Sign.mt, Lifeprint, Handspeak, Usely.ai, WeCapable, and ASL Meredith. The MediaPipe framework was used to obtain the hand landmarks in the case of both single-hand and dual-hand gestures, and the landmarks were then preprocessed to sync spatial features so that consistency across all scales and orientations was secured. A comparison of their performances showed that the MLP model outperformed the other machine learning approaches. Finally, model deployment with Android devices has been explored and benchmarked to help lay out a foundation for our future work, which includes extending gesture recognition to ISL, BSL, and integrating facial expression recognition for automatic text interpretation with offline inference. The suggested MLP model (1024,512) obtained a maximum of 99.3% classification accuracy with respect to 150 legal ASL signs.

Additionally, a separate lightweight prototype was developed to evaluate optimal deployment models for real-time offline inference with quantized and non-quantized models on mobile hardware (CPU, GPU, and TPU). This work focuses on isolated static sign recognition. Continuous video-based sign language modelling is left for future work. Although zero-padding is effective for dimensional consistency, future work will replace padding by employing a dataset that has explicit hand-presence to further reduce the possibility of shortcut learning.

Acknowledgement

This paper was supported by PM-USHA Research project Grant, Agreement Number: ROC No: SPMVV/UGC/F1/PM-USHA/2024.

References

[1] Sign Language, National Geographic Education, accessed January 10, 2026, https://education.nationalgeographic.org/resource/sign-language/.

[2] Seervai, H. M. Constitutional Law of India: A Critical Commentary with Supplement. 4th ed., Silver Jubilee Edition. New Delhi: Universal Law Publishing Co. Ltd. and Law & Justice Publishing Co., 2023.

[3] International Day of Sign Languages 2022, MIT Global Studies and Languages, accessed January 10, 2026, https://languages.mit.edu/international-day-of-sign-languages-2022.

[4] Dutta, Kusumika Krori, and Sunny Arokia Swamy Bellary. "Machine learning techniques for Indian sign language recognition." In 2017 international conference on current trends in computer, electrical, electronics and communication (CTCEEC), IEEE, 2017, 333-336.

[5] Cheok, Ming Jin, Zaid Omar, and Mohamed Hisham Jaward. "A Review Of Hand Gesture and Sign Language Recognition Techniques." International Journal of Machine Learning and Cybernetics 10, no. 1 (2019): 131-153.

[6] Rastgoo, Razieh, Kourosh Kiani, and Sergio Escalera. "Sign Language Recognition: A Deep Survey." Expert Systems with Applications 164 (2021): 113794.

[7] Chophuk, Ponlawat, and Kosin Chamnongthai. "Backhand-View-Based Continuous-Signed-Letter Recognition Using a Rewound Video Sequence and the Previous Signed-Letter Information." IEEE Access 9 (2021): 40187-40197.

[8] Buckley, Neil, Lewis Sherrett, and Emanuele Lindo Secco. "A CNN Sign Language Recognition System with Single & Double-Handed Gestures." In 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), IEEE, 2021, 1250-1253.

[9] AQ Mohammed, Adam, Yuan Gao, Zhilong Ji, Jiancheng Lv, M. D. Sajjatul Islam, and Yongsheng Sang. "Automatic 3D Skeleton-Based Dynamic Hand Gesture Recognition Using Multi-Layer Convolutional LSTM." In Proceedings of the 7th International Conference on Robotics and Artificial Intelligence, 2021, 8-14.

[10] Damdoo, Rina, and Ashutosh Gupta. "Gesture Controlled Interaction Using Hand Pose Model." International journal of health sciences I (2022): 10417-10427.

[11] Abdullahi, Sunusi Bala, and Kosin Chamnongthai. "American Sign Language Words Recognition Using Spatio-Temporal Prosodic and Angle Features: A Sequential Learning Approach." IEEE Access 10 (2022): 15911-15923.

[12] Adaloglou, Nikolas, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J. Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. "A Comprehensive Study on Deep Learning-Based Methods for Sign Language Recognition." IEEE transactions on multimedia 24 (2021): 1750-1762.

[13] Kothadiya, Deep, Chintan Bhatt, Krenil Sapariya, Kevin Patel, Ana-Belén Gil-González, and Juan M. Corchado. "Deepsign: Sign Language Detection and Recognition Using Deep Learning." Electronics 11, no. 11 (2022): 1780.

[14] Bose Duraimutharasan, Navaneetha Krishna, and Kumaravelu Sangeetha. "Machine Learning and Vision Based Techniques for Detecting and Recognizing Indian Sign Language." Revue d'Intelligence Artificielle 37, no. 5 (2023).

[15] Kumar, Chatikam Raj. "Machine Learning-Based Gesture Recognition for Communication with the Deaf and Dumb Prasanthi Yavanamandha1, Bodduru Keerthana2, Penmetsa Jahnavi3, Koduganti Venkata Rao4 and." Int. J. Exp. Res. Rev 34 (2023): 26-35.

[16] Kumar, Rupesh, Ashutosh Bajpai, and Ayush Sinha. "Mediapipe and CNNS For Real-Time Asl Gesture Recognition." arXiv preprint arXiv:2305.05296 (2023).

[17] Srivastava, Palak, Bramah Hazela, Shikha Singh, Pallavi Asthana, Deependra Pandey, Kamlesh Kumar Singh, and Vineet Singh. "Interpretation of Sign Language Using

Machine Learning." In 2024 Second International Conference Computational and Characterization Techniques in Engineering & Sciences (IC3TES), IEEE, 2024, 1-7.

[18] Josef, Antonio, and Gede Putra Kusuma. "Alphabet Recognition in Sign Language Using Deep Learning Algorithm with Bayesian Optimization." Revue d'Intelligence Artificielle 38, no. 3 (2024).

[19] Kumar, Mukesh, Chandradeep Bhatt, Shrestha Manori, Rahul Khati, Teekam Singh, and Taranath NL. "Sign Language Recognition by LSTM-Driven Speech Synthesis." In 2024 Second International Conference on Advances in Information Technology (ICAIT), vol. 1, IEEE, 2024, 1-6.

[20] Jana, Urmi, Subhashish Paul, and Dinabhandhu Bhandari. "Real-Time Caption Generation for the American Sign Language Using YOLO and LSTM." In 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), IEEE, 2024, 1-4.

[21] Fertl, Elfi, Encarnación Castillo, Georg Stettinger, Manuel P. Cuéllar, and Diego P. Morales. "Hand Gesture Recognition on Edge Devices: Sensor Technologies, Algorithms, and Processing Hardware." Sensors (Basel, Switzerland) 25, no. 6 (2025): 1687.

[22] Riya Awalkar, Aditi Sah, Renuka Barahate, Yash Kharche, Ashwini Magar (2025), Silent Expressions: Two-Handed Indian Sign Language Recognition Using MediaPipe and Machine Learning. International Journal of Innovative Science and Research Technology (IJISRT) IJISRT25MAR598, 587-595.

[23] Donepudi, Swapna, K. N. Divya, Reemasen Tungala, and Sirisha Thammisetty. "A Sign Language Recognition System for Understanding Actions of Speech-Impaired-Individuals." In 2025 7th International Conference on Intelligent Sustainable Systems (ICISS), IEEE, 2025, 1303-1310.