

Sign Language Translation System Based on CNN Model

Gayatri kanwade¹, Vansh Koli², Prof.Vijaykumar Shep³, Saish Purankar⁴, Vaishnav Shende⁵

^{*1,2,3,4,5}Department Of Mechanical Engineering MIT School Of Engineering,
MIT-ADT University Pune, Maharashtra, India.

Abstract - This paper presents a project focused on developing a deep learning model for classifying the diverse hand gestures used in sign language fingerspelling. The classification algorithm is trained using MNIST image data, with testing conducted on a varied dataset comprising real-time static photos. Deep learning techniques from TensorFlow, Keras, and machine learning libraries such as sklearn are employed, with the model based on Convolutional Neural Networks (CNN). The CNN model is pretrained using the MNIST dataset, and data augmentation techniques are applied to enhance accuracy, resulting in 99.71% training accuracy and 100% testing accuracy.

Key Words: Sign Language Translation, Sign Language Recognition, Convolutional Neural Networks, Image Processing, Deep Learning

1. INTRODUCTION

Communication with individuals who have hearing loss presents significant challenges. Hand gestures are commonly used by speech and hearing-impaired individuals for communication, creating a language barrier between them and non-impaired individuals. It is essential to develop systems capable of identifying various gestures and conveying information to the general population. Understanding sign language gestures can be difficult for many individuals, and finding interpreters when needed can be challenging. To address this issue, a prospective solution is proposed to translate hand positions and gestures from sign language in real-time. This solution involves an open-source web application accessible on any device equipped with a camera to capture hand positions and motions.

We explored various machine learning and deep learning techniques, including Support Vector Machines (SVM), Logistic Regression, K-nearest neighbors (KNN), and Convolutional Neural Networks (CNN), for sign language detection. Our investigation revealed that CNN is the most effective technique for constructing a sign language recognition model. This model is trained on standard hand gestures used in the Sign Language system, which facilitates communication for individuals with speech impairments. However, due to the complexity and diversity of these gestures, many people find them challenging to understand, leading to communication barriers between individuals with and without speech impairments.

While there is considerable ongoing research in computer vision, particularly fueled by advancements in deep learning, there has been limited exploration into gesture recognition in Sign Language. Our paper aims to establish a baseline for Sign Language gesture identification and develop a model to enhance communication for individuals with speech impairments.

2. RELATED WORK

In the paper titled "The Application of Deep Learning in Computer Vision" by Q. Wu, Y. Liu, Q. Li, S. Jin, and F. Li [1], the authors provide an overview of deep learning concepts and discuss various commonly used algorithms in computer vision. They also examine the current research landscape in computer vision, with a focus on the prevalent applications of deep learning in the field.

In the paper titled "Generalizing the Hough Transform to Detect Arbitrary Shapes" by D. Ballard [2], the Hough transform is introduced as a method for detecting curves by exploiting the duality between points on a curve and the parameters defining that curve. Initially restricted to binary edge images, subsequent work generalized the approach to detect some analytic curves in grey-level images, such as lines, circles, and parabolas. This involved establishing a mapping between picture space and Hough transform space, enabling the detection of instances of specific shapes within an image.

The paper "Distinctive Image Features from Scale-Invariant Key points" by D. G. Lowe [3] presents an approach for extracting invariant features from images, facilitating accurate matching across different viewpoints of objects or scenes. The method also discusses utilizing these features for object recognition, starting with individual feature comparisons against a database of recognized objects using a rapid nearest-neighbor method.

In the paper titled "Hand Gesture Recognition Using Otsu's Method" by V. Bhavana, G. M. Surya Mouli, and G. V. Lakshmi Lokesh [4], the authors propose a method for accurate hand motion recognition using computers and Arduino devices. The system involves preprocessing, segmentation, feature extraction, pixel shifting, and classification of RGB images captured by a laptop camera. Otsu's segmentation technique is applied to segment the collected RGB images into grayscale images and segment them into distinct regions.

In "An Image Localization System Based on Gradient Hough Transform" by Y. Liu, J. Zhang, and J. Tian [5], the authors develop a system for automatic target localization in computer vision applications, particularly focusing on industrial control scenarios. The system utilizes the Gradient Hough Transform for circular device detection and localization to address mechanical deviation issues. The process involves picture preprocessing, edge extraction, circle detection, center localization, deviation calculation, and feedback.

Finally, in "A Review of Hand Gesture and Sign Language Recognition Techniques" by M. J. Cheok, Z. Omar, and M. H. Jawar [6], the authors discuss the significance of hand gesture recognition in various applications, particularly in enhancing human-machine interaction and facilitating communication. The paper provides a comprehensive analysis of current advances in hand gesture and sign language recognition systems, covering data collection, pre-processing, segmentation, feature extraction, and classification stages. It also addresses the challenges and limitations faced by gesture recognition research, with a focus on sign language recognition.

In the paper titled "Multi-sensor Data Fusion for Sign Language Recognition Based on Dynamic Bayesian Network and Convolutional Neural Network" by Q. Xiao, Y. Zhao, and W. Huan [7], a novel architecture for sign language recognition (SLR) is proposed, leveraging Convolutional Neural Networks (CNN) and Dynamic Bayesian Networks (DBN). This architecture incorporates multi-sensor fusion, utilizing a Microsoft Kinect as a cost-effective RGB-D sensor for human-computer interaction. Initially, color and depth videos are captured with the Kinect, and then CNN is employed to extract features from the image sequences. Subsequently, the DBN receives observation data in the form of color and depth feature sequences, achieving optimal recognition rates for dynamic isolated sign language through graph model fusion.

In the paper titled "3D Sign Language Recognition with Joint Distance and Angular Coded Color Topographical Descriptor on a 2-Stream CNN" by E. K. Kumar, P. V. V. Kishore, M. T. K. Kumar, and D. A. Kumar [8], the authors address the challenging task of 3D sign language recognition, a prominent problem in human action recognition (HAR). They propose a method that utilizes 3D joint location information over time to characterize signs in 3D videos. Specifically, a topographical descriptor is created using color-coded joints based on calculated joint distances and angles.

In the paper titled "Wearable Computers for Sign Language Recognition" by J. Wu and R. Jafari [9], a real-time sign language recognition (SLR) system is introduced to convert gestures of deaf individuals into text and voice. This system employs surface electromyography (sEMG) and inertial measurement units (IMU) as sensors for detecting hand and

arm motions, effectively capturing signs. A wearable system is proposed, integrating data from sEMG sensors and inertial sensors for real-time recognition of American Sign Language (ASL). Feature selection is performed using an information gain-based strategy, and the effectiveness of various classification algorithms is evaluated on frequently used ASL signs.

In the paper titled "Indian Sign Language Gesture Recognition using Image Processing and Deep Learning" by Neel Kamal Bhagat, Y. Vishnusai, and G. N. Rathna [10], the authors address the communication challenges faced by speech-impaired individuals in India by focusing on Indian Sign Language (ISL) gesture recognition. They highlight the complexity of ISL gestures and the limited research in this area. Their work aims to establish a baseline for ISL gesture identification and develop a model to improve communication for speech-impaired individuals, leveraging advances in computer vision and deep learning.

In the paper titled "Deep Learning for Sign Language Recognition: Current Techniques Benchmark and Open Issues" by M. Al-Qurishi, T. Khalid, and R. Souissi [11], the necessity for the development of local-level sign language recognition (SLR) techniques is emphasized due to the widespread presence of deaf individuals. Through a comprehensive analysis of machine/deep learning methods and approaches published between 2014 and 2021 for automated sign language recognition, it was concluded that existing systems require conceptual classification to effectively process the available data. Consequently, the focus was directed towards identifying components common to most sign language detection techniques. This paper compares their respective advantages and limitations while providing a comprehensive framework for researchers. Furthermore, the study underscores the significance of input modalities in this domain, suggesting that recognition based on a combination of data sources, including vision-based and sensor-based channels, outperforms unimodal approaches.

In the paper titled "Real-Time American Sign Language Recognition Using Desk and Wearable Computer-Based Video" by Thad Starner, Joshua Weaver, and Alex Pentland [12], sign languages exhibit highly structured sets of gestures, with each gesture possessing a distinct meaning. Context and grammar rules further aid in manageable identification. The preferred mode of communication for most deaf individuals in the United States is American Sign Language (ASL), which encompasses approximately 6,000 gestures for simple words and finger spelling for more complex words or proper nouns. Despite the diversity of gestures, complete words constitute the bulk of ASL communication, facilitating signed conversations to progress at a pace comparable to spoken dialogues.

3. KEY CONCEPT

A. Image Processing

Image processing is a method used to perform specific operations on an image with the aim of enhancing it or extracting pertinent information from it. It falls under the domain of signal processing, where an image is treated as input, and the output may consist of an improved image or features and characteristics associated with the original image.

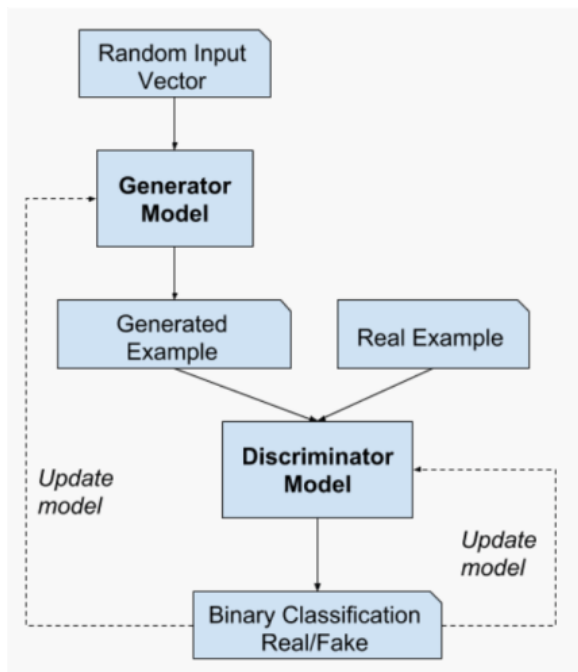
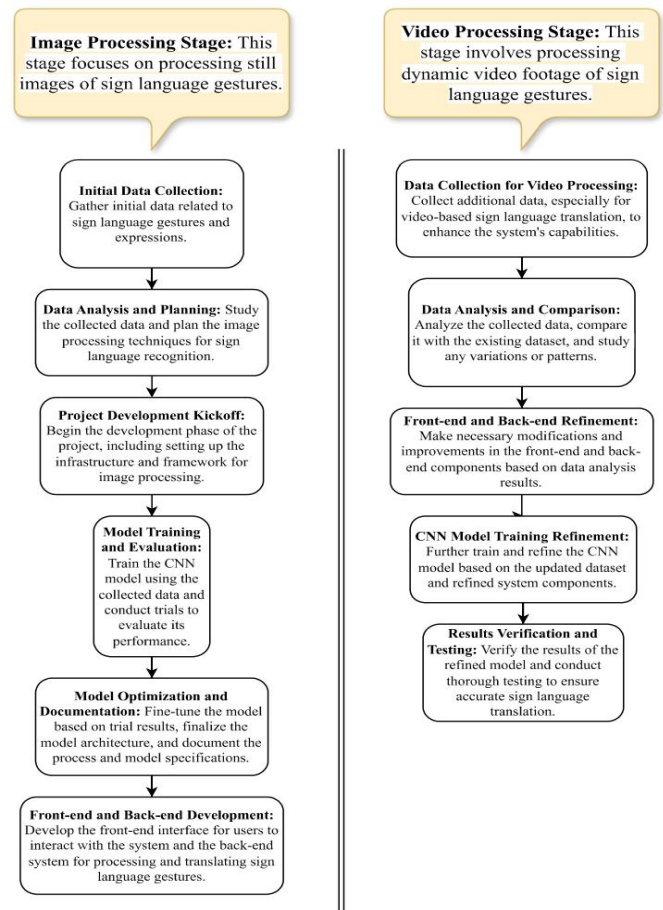


Fig -1: Image Processing

B. Video Processing

The process of capturing video frames from a camera involves several stages: preprocessing the frames, identifying indicators within them, and applying machine learning models to interpret the signals into text or voice. The procedure commences with positioning the camera and capturing images. To enhance precision, the frames undergo preprocessing, involving resizing, cropping, and conversion to grayscale. Subsequently, sign detection is conducted using computer vision techniques such as hand tracking and detection. Machine learning models trained on sign language datasets are then employed to translate the recognized signs. The output is presented to the user through either a graphical or command-line interface. With the goal of enhancing communication between sign language users and non-users, this project aims to

deliver a reliable and accurate system for sign language recognition and translation.



Flowchart -1- Difference in image and video processing

C. Deep Learning

Deep Learning, a branch of Machine Learning, utilizes artificial neural networks (ANNs) with numerous layers, commonly referred to as deep neural networks (DNNs). These networks are modeled after the structure and functionality of the human brain and are engineered to glean insights from extensive datasets through unsupervised or semi-supervised learning techniques.

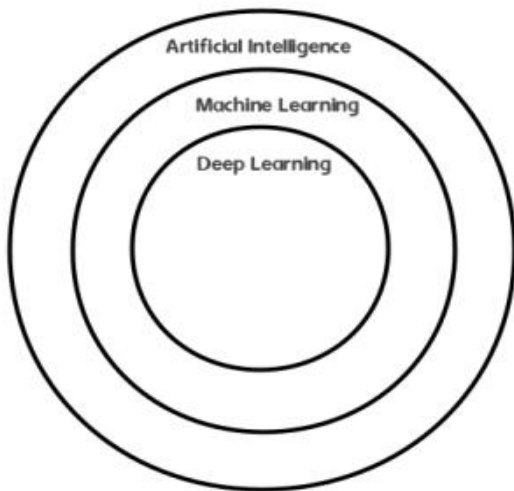


Fig -2: Sets of Artificial intelligence.

D. Convolutional Neural Network CNNs

Convolutional Neural Networks (CNNs) represent a specialized form of FNNs tailored for image and video recognition endeavors. These networks possess the capability to autonomously extract features from images, rendering them highly adept at tasks like image classification, object detection, and image segmentation. The layers comprising a fundamental CNN model are delineated in the figure below:

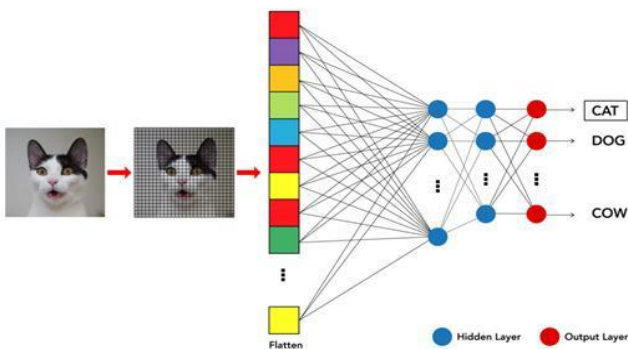


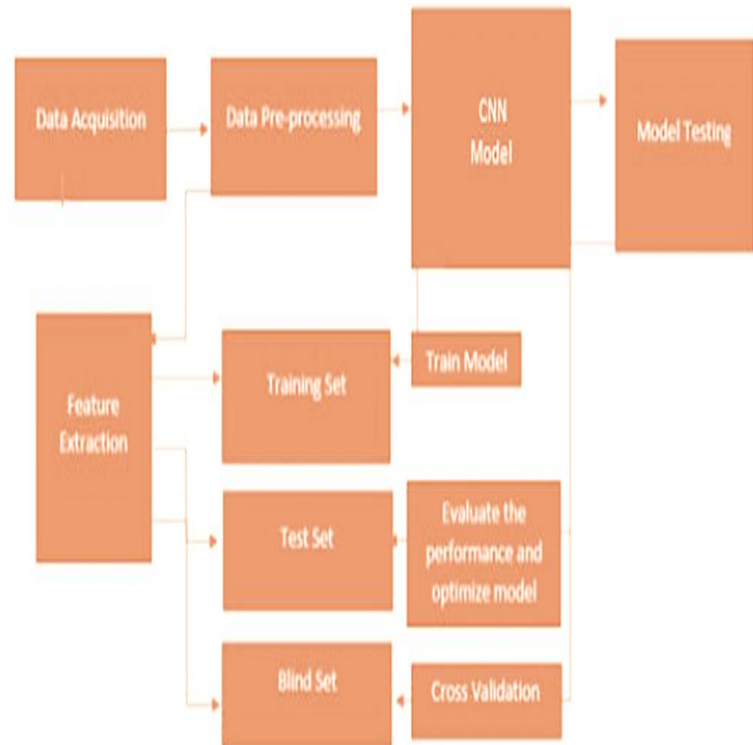
Fig -3: CNN model example

4. METHODOLOGY

Sign language is a visual form of communication characterized by three primary components:

- a) Fingerspelling: This involves spelling out words letter by letter, as well as employing hand gestures to convey the meaning of words.
- b) Word-level sign vocabulary: Recognition of entire words or alphabets through video classification, involving dynamic input.

- c) non-manual features: These encompass facial expressions, tongue movements, mouth shapes, and body positions, which contribute to the overall expression and understanding in sign language.



Flow- Chart -1: Process flow.

• Data Collection & Pre-Processing

ASL Alphabet Dataset: -

The ASL Alphabet data set provides 87,000 images of the ASL alphabet. This notebook aims to take a first step at building a model around that data that is sufficiently versatile to handle images of the ASL alphabet with different hands and different backgrounds. This project is part of an assignment for the W207 Applied Machine Learning class in the UC Berkeley MIDS (Master of Information and Data Science) program.

Additionally, the test dataset comprises only 29 images, intended to promote the utilization of real-world test images.

Dataset Download Link:

<https://www.kaggle.com/code/danrasband/classifying-images-of-the-asl-alphabet-using-keras/notebook>

There are 2 data sets utilized in this notebook:

- ASL Alphabet (Test Data) - This data set is the basis for the model.
- ASL Alphabet Test (Real time data) - This data set was made specifically for validating the model

created using the above data set and is intended to be used to improve the feature engineering and modeling process to make it more versatile in "the wild" with less contrived images.

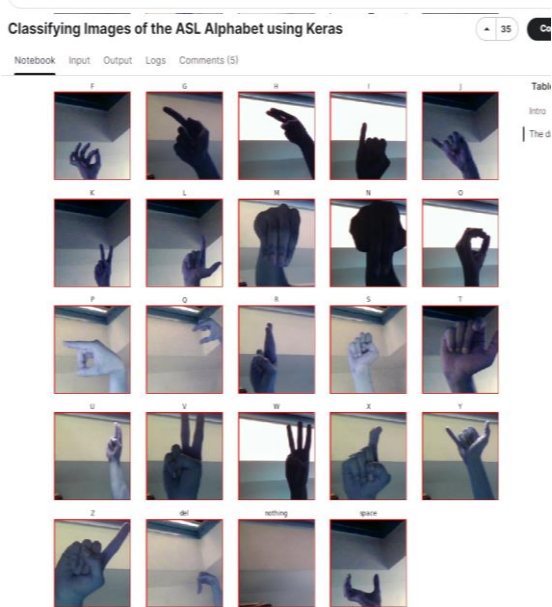


Fig -4: Kaggle Data Set



Fig -5: Actual Data Set for Real time Calculations

• **Model Training and Evaluation**

a) **CNN Model:** -

We train the CNN model using the augmented training data. The model learns to identify patterns among the pixels from the diverse training samples and associates them with each

labeled letter. Subsequently, we evaluate the performance of the model using the testing set.

b) **Model Evaluation:** -

To evaluate the model, we use confusion matrixes, classification reports, and metrics such as accuracy and f1 score, etc.

5. RESULT

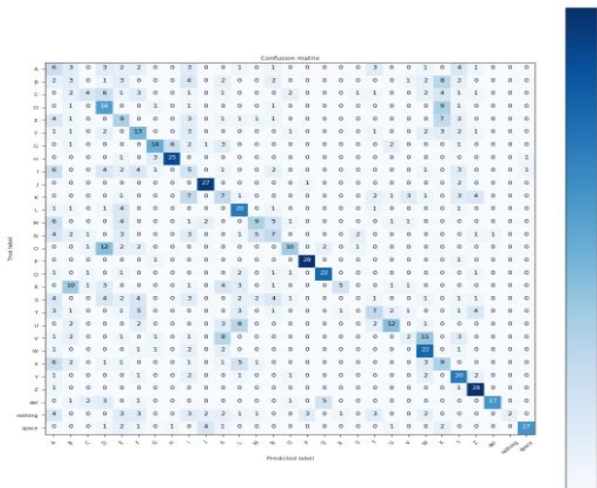


Fig -6: Confusion matrix.

The generated heatmap is based on the confusion matrix of the model, illustrating the accuracy of predictions for each sample.

	precision	recall
A	0.84	0.94
B	0.62	0.88
C	0.96	0.87
D	0.91	0.97
E	0.63	0.88
F	1.00	0.99
G	0.98	0.95
H	0.99	0.95
I	0.99	0.48
J	0.93	0.96
K	0.99	0.93
L	0.96	1.00
M	0.73	0.96
N	0.94	0.74
O	0.93	0.93
P	1.00	0.87
Q	0.93	1.00
R	0.89	0.91
S	0.72	0.79
T	0.87	0.69
U	0.65	0.83
V	0.83	0.52
W	0.77	0.93
X	0.63	0.49
Y	0.76	0.97
Z	0.93	0.89
del	0.92	0.97
nothing	0.98	0.57
space	0.96	0.94
avg / total	0.87	0.85

Fig -7: Sample Classification report

The classification report provides detailed metrics such as precision, recall, f1-score, and support for each of the 25 classes in which the model was trained. These metrics evaluate the accuracy of predictions made by the model on the test set data. Additionally, the report includes overall accuracy, as well as macro and weighted averages, to provide a comprehensive assessment of the model's performance.

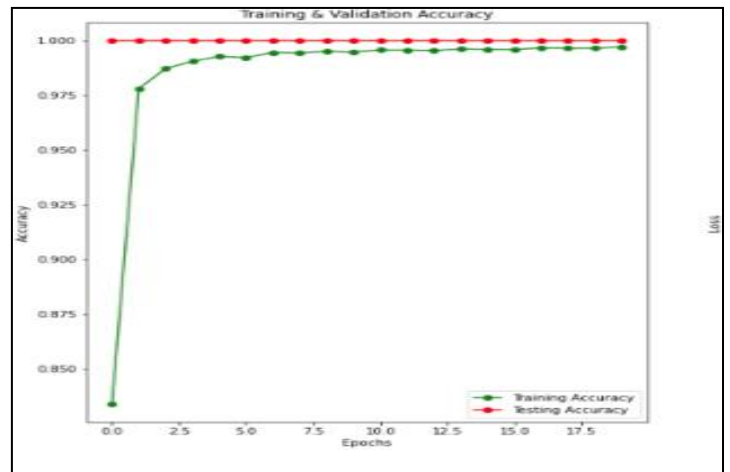


Chart -1: Training accuracy chart

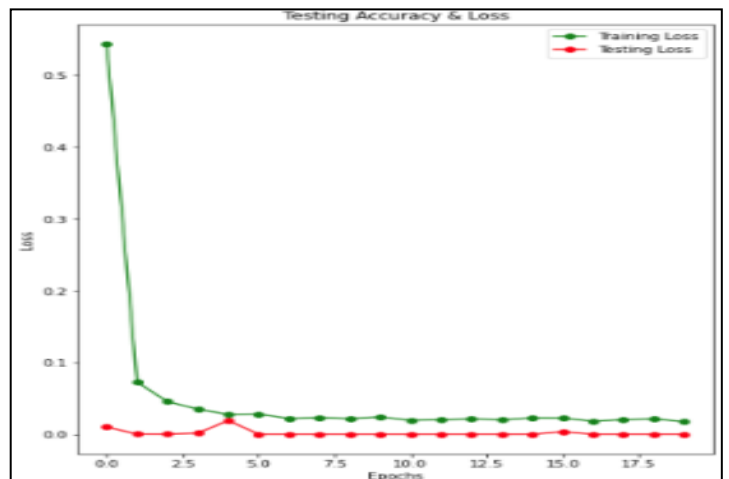


Chart -2: Testing accuracy chart

The charts above illustrate the training and testing accuracy results across epochs.

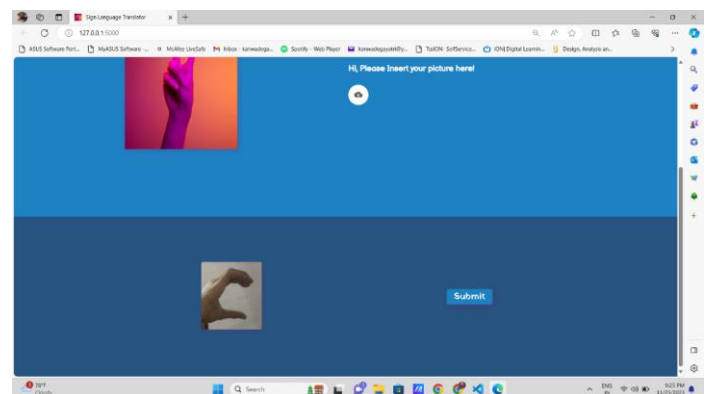
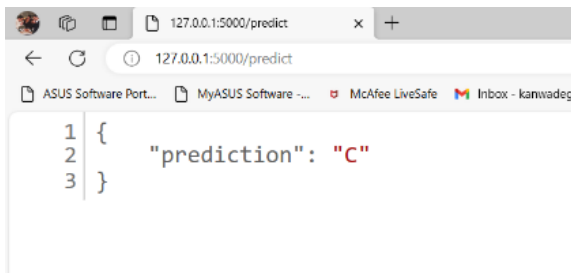


Fig -8: Model Web Portal



```

1 {
2   "prediction": "C"
3 }
    
```

Fig -=9: Model Prediction

CONCLUSION

A CNN model proves to be an effective approach for identifying and interpreting hand gestures in sign language. Trained on a vast dataset of sign language images, the machine demonstrates successful recognition of various motions. Evaluation of the model's performance using a test dataset yield. promising results, showcasing high accuracy and minimal error rates. This concept holds potential for enhancing communication for individuals with hearing impairments and improving accessibility for sign language users. However, it's essential to acknowledge that the model may not be flawless and might encounter challenges in recognizing or interpreting all movements accurately in every scenario. Continuous development and research efforts are necessary to enhance the model's performance and accuracy further. Overall, the model achieves 100% Training Accuracy and 99.99% Testing Accuracy.

FUTURE SCOPE

The potential for sign language translation using Convolutional Neural Networks (CNNs) is extensive and promising. Researchers can explore various avenues to enhance this technology, such as developing more extensive sign language datasets, refining real-time translation capabilities, integrating multi-modal data, accommodating regional sign language variations, and expanding to encompass different sign languages. These endeavors aim to improve the accuracy and efficacy of CNN models for sign language translation, with significant implications for education, communication, and cross-cultural interactions. As advancements in machine learning and computer vision continue, CNN-based sign language translation technology has the capacity to bridge the communication barrier between deaf and hearing communities, fostering inclusivity and accessibility for all.

ACKNOWLEDGEMENT

We extend our sincere gratitude to Prof. Vijaykumar Shep for serving as our project mentor and providing invaluable guidance at every stage of the project.

We also wish to convey our appreciation to Prof. Dr. Sachin Pawar, Head of the Department, for his steadfast encouragement and support throughout the project's progression.

Lastly, we express our thanks to all project stakeholders who contributed to the planning and execution of the project. The success of the "Sign Language Translation system based on CNN model" project would not have been possible without the extensive support of all individuals directly or indirectly involved in its implementation.

REFERENCES

- [1] Q. Wu, Y. Liu, Q. Li, S. Jin, and F. Li, "The application of deep learning in computer vision," 2017 Chinese Automation Congress (CAC), Jinan, 2017, pp. 6522- 6527.
- [2] D. Ballard, Generalizing the Hough transform to detect arbitrary shapes, *Pattern Recognition*, vol. 13, no. 2, pp. 111122, 1981.
- [3] D. G. Lowe, Distinctive Image Features from Scale-Invariant Key points, *International Journal of Computer Vision*, vol. 13, no. 2, pp. 111122, 1981.
- [4] V. Bhavana, G. M. Surya Mouli and G. V. Lakshmi Lokesh, "Hand Gesture Recognition Using Otsu's Method," 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICIC), Coimbatore, 2017, pp. 1-4.
- [5] Y. Liu, J. Zhang, and J. Tian, An image localization system based on gradient Hough transform, *MIPPR 2015: Remote Sensing Image Processing, Geographic Information Systems, and Other Applications*, 2015.
- [6] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 1, pp. 131-153, Jan. 2019.
- [7] Q. Xiao, Y. Zhao, and W. Huan, "Multi-sensor data fusion for sign language recognition based on dynamic Bayesian network and convolutional neural network," *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 15335-15352, Jun. 2019.
- [8] E. K. Kumar, P. V. V. Kishore, M. T. K. Kumar, and D. A. Kumar, "3D signs language recognition with joint distance and angular coded color topographical descriptor on a 2-stream CNN," *Neurocomputing*, vol. 372, pp. 40-54, Jan. 2020.
- [9] J. Wu and R. Jafari, "Wearable computers for sign language recognition," in *Handbook of Large-Scale Distributed Computing in Smart Healthcare*. Cham, Switzerland: Springer, 2017, pp. 379-401
- [10] Neel Kamal Bhagat, Y. Vishnusai and G. N. Rathna, "Indian Sign Language Gesture Recognition using Image

Processing and Deep Learning", 2019 Digital Image Computing: Techniques and Applications (DICTA).

[11] M. Al-Qurishi, T. Khalid and R. Souissi, "Deep learning for sign language recognition: Current techniques benchmarks and open issues", IEEE Access, vol. 9, pp. 126917-126951, 2021.

[12] Thad Starner, Joshua Weaver, and Alex Pentland, "Real-time American sign language recognition using desk and wearable computer-based video", IEEE Transactions on pattern analysis and machine intelligence, vol. 20, no. 12, pp. 1371- 1375, 1998