# Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training

Author Ms. Shyla N, Suprith Satish, Vijay Payyavula, Vishal Veeraraj Shetty, Yash S Sindhe
*Department of Information Science & Engineering*
*RNS Institute of Technology*
*Bangalore, India.*

***Abstract***— American Sign Language (ASL) is a language used by people who have hearing and speech difficulties. It's necessary for their expression and conversation. The complexities of the language, as well as significant interclass differences, make it challenging to recognize ASL using computer vision. Convolutional neural networks (CNNs) are a promising solution to this problem. A CNN is a deep learning system used for image recognition. They have been used successfully in a variety of applications, including gesture, face, and object recognition. In the context of interpreting ASL gestures, CNNs can be trained to detect the hand motions that correspond to each letter of the ASL alphabet. This requires giving the model a large variety of images of ASL motions to teach it the patterns and traits that distinguish each letter. The use of CNNs to recognize ASL can dramatically increase accessibility and communication opportunities for people with hearing and speaking disabilities. By recognizing ASL motions, computers can convert them into written or spoken words, allowing for seamless communication between people with and without disabilities. Furthermore, this technology might be integrated into a multitude of devices and platforms, such as laptops, tablets, and smartphones, making it easily available to individuals in need. Finally, the ability of CNNs to recognize ASL gestures has the potential to dramatically improve the quality of life for persons with hearing and speech disabilities.

**Keywords; NB algorithm, Dataset.**

## I. INTRODUCTION

The Sign Language Alphabets are constructed by using hand and facial motions to express the thoughts of the deaf and hard of hearing population to those who do not understand sign language. This causes a major communication gap between the general public and persons with disabilities. There have been technological breakthroughs to assist speech and hearing challenged people because it is impossible for them to seek assistance all of the time and assist them in their daily activities as a translator. An innovative approach can employ technology to turn sign gestures into comprehensible words for the general public.

This technique will bridge the communication gap and facilitate communication between the disabled and the general public. Indian Sign Language, American Sign Language (ASL), Italian Sign Language, and many other sign languages are used around the world to help deaf and speech impaired people communicate. Sign Language is used as the major form of communication by thousands and millions of people worldwide, including millions in India. American Sign Language is the most extensively used sign language, ranking fourth in North America. It may surprise you to learn that, in addition to the United States, ASL is used in more than 30 additional countries where English is the primary means of communication. ASL is used as the principal form of communication by around a million people in the United States and other parts of the world [1]. ASL is a wide and complicated language that is developed by utilizing hand, finger activities, and facial gestures to represent the thoughts of persons who are deaf or hard of hearing. ASL Language is recognized as a suitable and original language, with more variations than any other language.

Italian and German are two examples. Sign language is a great piece of interaction art that benefits a large population of deaf and hard of hearing people. The existence of ASL encourages happiness and hope among persons who are visually challenged, and its broad influence is amazing and eye-catching [2]. ASL is a set of 26 gesture signs known colloquially as the American Manual Alphabet that can be used to convey various words from the English Dictionary. ASL uses 19 various hand forms to communicate, which contributes to the formation of 26 American manual alphabets. Because there are fewer hand forms accessible, some hand shapes express distinct alphabets depending on their orientation. Letters such as 'K' and 'P' are examples. Hand gestures can also be used to convey the numbers '0' through '9'. Despite the fact that there are no hand signals for specific nouns and concepts [3], different facial and hand gesture indicators exist to represent various English words. Figure 1 depicts ASL's set of 10 numbers ranging from 0 to 9 and 26 alphabet gesture signs of English alphabets from A to Z. ASL, or American Sign Language, is a natural language with many of the same characteristics as spoken language. As previously stated, ASL is utilized as a means of conversation by employing hand and facial movements. It is one of the primary ways of communication for most North Americans with speech and hearing difficulties, and it is also utilized by non-disabled people. If one individual is unfamiliar with the language, sign language recognition will facilitate communication.

### Issue Statement

The detection of the hand sign is a difficult operation. It takes time as well. As a result, we employ Machine Learning to detect spam.

### Existing System

Sign language is the most natural and effective means for deaf and hearing individuals to communicate with one another. Due to challenges in hand segmentation and appearance variations among signers, American Sign Language (ASL) alphabet recognition (i.e. fingerspelling) with a marker-less visual sensor is a difficult process. Existing color-based sign language recognition systems face numerous obstacles, including complicated background, hand segmentation, high inter-class and intra-class variability, and a deep learning architecture based on the Principal Component Analysis Network (PCANet). On this detection, two learning algorithms for the PCANet model are given.

### Disadvantages:

Existing color-based sign language recognition systems require a significant amount of time and it uses complex methodologies and has a convoluted background, such as hand segmentation.ac

### Proposed System

We are working on a machine learning project in the proposed system. The sign language is being detected. We are predicting sign language using the CNN algorithm. Implementing the CNN model in a real-time system capable of recognizing user-inputted gestures. A pass input image will be used in the system to record the user's hand motions and feed them to the CNN model for recognition.

### Advantages:

Using a CNN algorithm to recognize hand motions and translate them into sign language can yield high accuracy. This can result in a more reliable and efficient system capable of providing precise translations of sign language motions. The suggested system can save time by translating sign language motions in real time without the need for a human translator. This can save time when communicating between deaf and hearing people, as well as in circumstances requiring sign language interpretation, such as in classrooms, medical settings, and public areas. Implementing the CNN model in a real-time system can result in a speedier recognition process. This can be useful in instances where immediate sign language translation is required, such as in an emergency or in a fast-paced atmosphere. Furthermore, as the system learns and adapts to new data, the usage of a machine learning project can give faster and more efficient recognition of hand movements over time.

## II. RELATED WORK

Zhi-hua Chen et al., “Real-Time Hand Gesture Recognition Using Finger Segmentation”.[1]

Hand gesture recognition is critical for human-computer interaction. In this paper, we describe a revolutionary real-time approach for hand gesture identification. The hand region is extracted from the backdrop in our framework using the background subtraction method. The palm and fingers are then split to detect and recognise the fingers. Finally, a rule classifier is used to predict the labels of hand movements. Experiments on a data set of 1300 photos reveal that our method performs well and is highly efficient. Furthermore, on another data set of hand movements, our method outperforms a state-of-the-art method.

A. Kumar et al., “Dog Breed Classifier for Facial Recognition Using Convolutional Neural Networks”.[2]

This study was about dog breed classification. Classifying dog breeds is a difficult task for a deep convolutional neural network. A set of sample photos of a breed of canines and humans is used to classify and learn the breed's characteristics. Image processing converts the images to a single dimension label. The photos of humans and dogs are used for breed categorization in order to determine the existing percentage of features in humans and dogs. This study employed principal component analysis to combine the most comparable features into one group, making it easier to investigate the features in deep neural networks. Furthermore, the facial features are saved as vectors. This vector will be compared with each feature of the dog in the database to yield the most efficient outcome. 13233 human photos and 8351 dog photographs are considered in the suggested experiment. The photos under test are categorised as a breed based on the weight difference between test and train images. This study is based on research that uses CNN to classify different dog breeds. If a dog image is provided, the algorithm will work to determine the breed of dog and feature similarity in the breed. If a human image is provided, the algorithm calculates the facial features that exist in a dog of human and vice versa.

Asif Karim et al., “User-independent American Sign Language Alphabet Recognition Based on Depth Image And PCANet Features”. [3]

Sign language is the most natural and effective means for deaf and hearing people to communicate. Due to challenges in hand segmentation and appearance variations among signers, recognising American Sign Language (ASL) alphabets (i.e. fingerspelling) using marker-less vision sensors is a difficult process.

Existing color-based sign language identification systems face numerous obstacles, including complicated background, hand segmentation, and high inter- and intra-class variances. We present a new user-independent recognition method for

the American sign language alphabet based on depth photos collected by the low-cost Microsoft Kinect depth sensor in this work. Because depth information is resistant to fluctuations in illumination and background, using it instead of color images solves several problems.

A hand region can be segregated using a simple preprocessing approach applied to a depth image. Instead of the traditional hand-crafted feature extraction approaches, feature learning using convolutional neural network topologies is used. A simple unsupervised Principal Component Analysis Network (PCANet) deep learning architecture is used to successfully learn local characteristics retrieved from the segmented hand.

Two learning methodologies for the PCANet model are proposed: train a single PCANet model from samples of all users and train a separate PCANet model for each user. Linear Support Vector Machine (SVM) classifiers are then used to recognise the retrieved features. The suggested method's performance is evaluated using a public dataset of real depth photographs obtained by diverse users.

The suggested method outperforms state-of-the-art recognition accuracy utilising a leave-one-out evaluation strategy, according to experimental results.

Walaa Aly et al., "User-independent American Sign Language Alphabet Recognition Based On Depth Image And PCANet Features".[4]

Sign language is the most natural and effective means for deaf and hearing people to communicate. Due to challenges in hand segmentation and appearance variations among signers, recognising American Sign Language (ASL) alphabets (i.e. fingerspelling) using marker-less vision sensors is a difficult process. Existing color-based sign language identification systems face numerous obstacles, including complicated background, hand segmentation, and high inter- and intra-class variances. We present a new user-independent recognition method for the American sign language alphabet based on depth photos collected by the low-cost Microsoft Kinect depth sensor in this work. Because depth information is resistant to fluctuations in illumination and background, using it instead of colour images solves several problems.

A hand region can be segregated using a simple preprocessing approach applied to a depth image. Instead of the traditional hand-crafted feature extraction approaches, feature learning using convolutional neural network topologies is used. A simple unsupervised Principal Component Analysis Network (PCANet) deep learning architecture is used to successfully learn local characteristics retrieved from the segmented hand. Two learning methodologies for the PCANet model are proposed: train a single PCANet model from samples of all users and train a separate PCANet model for each user.

The collected features are then recognized using linear Support Vector Machine (SVM) classifiers. The suggested method's performance is evaluated using a public dataset of real depth photographs obtained by diverse users. The experimental results reveal that the proposed method outperforms state-of-the-art recognition accuracy when utilizing the leave-one-out evaluation strategy.

Dr. Thamaraiselvi et al., "Sign Language Recognition Using CNN".[5]

Sign language is a linguistically full, mostly visual-spatial language. It is, in general, the primary language and consequently the major mode of communication for deaf people. Many individuals are aware of sign language. It is also not a universal language, contrary to common perception. As a result, the communication gap between the Deaf community and hence the hail maturity grows. Written communication is time consuming, simple, and preferred only when people are immobile. While walking or moving, written communication is frequently awkward. Furthermore, the Deaf community has a lesser level of proficiency in writing speeches. The primary goal of a hand sign recognition system is to establish a trade between mortal and CNN classifiers in which the honored signs are commonly employed for expressing useful information or can be used to provide input to a machine without touching physical clods and dials on that machine. In our paper, we present a model created with Convolutional Neural Networks.

## III. METHODOLOGY

The System architecture is shown below.

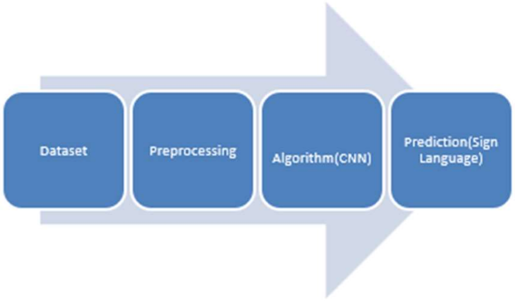


**Figure 1: System Architecture**

The architecture has 4 main steps:

1. Dataset
2. Data Pre-processing
3. Algorithm
4. Prediction

### Dataset

The collection of photos in sign language Typically, a dataset contains a huge number of photographs of people making hand motions to represent various signs in sign language. Machine learning models can be trained using these photos to recognize and interpret sign language motions into text or voice.

Images of hands and arms in various positions and orientations, shot from various perspectives and distances, may be included in the dataset. Each image can be labelled with the sign language word or phrase that it depicts. The collection may additionally include photographs of background scenes, lighting settings, and other environmental elements that may influence sign language gesture recognition.

### Data Preprocessing

Data preprocessing is the process of preparing raw data for use with a machine learning model. It is the first and most important stage in developing a machine learning model. When developing a machine learning project, we do not always come across clean and prepared data. And, before doing any operation on data, it must be cleaned and formatted.

As a result, we apply the data preprocessing task for this implementation.

### Algorithm (CNN)

Convolutional Neural Networks are a complicated neural network chain that works to extract visual features from a trained dataset and classify them to get the desired output. It trains the neural networks by converting the dataset images to numerical values. The fundamental advantage of CNN over its predecessors is that it automatically discovers significant elements without the need for human intervention. ConvNets outperform machine learning methods in terms of power and efficiency.

Based on their categorized properties, these numerical values are subsequently placed in numerical arrays. These arrays are then placed in various network nodes and iterated through several times based on the input provided.
The CNN models are used for geographical classification in multiple companies which require data to be classified in a quick and secure way it almost acts like a filter removing dust and separates the features.

### Prediction

The model takes input photos of hand movements, collects their attributes, and predicts the matching sign language word or phrase during the prediction phase. Images can be taken in real time with a camera or sensor, or they can be pre-recorded and analyzed later.

The machine learning model can be fine-tuned on specific sign languages or dialects to increase prediction accuracy, or it can be trained on a broader range of sign languages to provide multilingual translation.

The model can also be optimized for multiple types of sign language, such as one-handed or two-handed signing, as well as varied levels of sign language complexity and nuance.

## IV. ALGORITHMS

### Convolutional Neural Network (CNN):

Convolutional Neural Networks, often known as CovNets, are neural networks that share parameters. Assume you have an image. It can be represented as a cuboid with length, breadth (picture size), and height (since images often contain red, green, and blue channels).

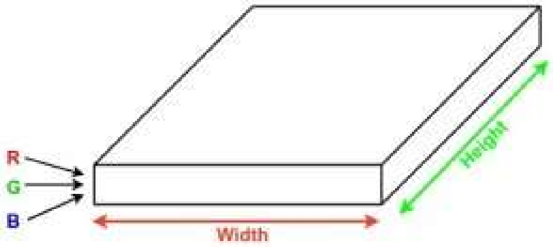


**Figure 2: Representation of image in cuboid form**

Consider taking a small area of this image and running a little neural network with, say, k outputs on it, and representing it vertically. Slide that neural network across the entire image, and we'll obtain a new image with varied width, height, and depth. Instead of only the R, G, and B channels, we now have more channels with a smaller width and height. Convolution is the name of his operation. If the patch size is the same as the image size, the neural network is a normal neural network. We have fewer weights as a result of this tiny patch.

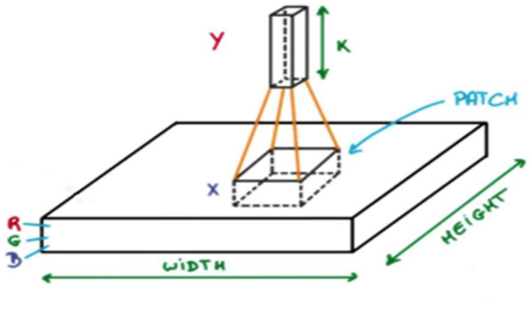


**Figure 3: Tiny area in the image that is analyzed by CNN.**

Let us now discuss some mathematics that is involved in the convolution process.

- Convolution layers are made up of a collection of learnable filters (seen as a patch in the above image). Every filter has a tiny width and height, as well as the same depth as the input volume (3 if the input layer is an image).
- For example, suppose we need to conduct convolution on a 34x34x3 image. Filter sizes can be axax3, where 'a' can be 3, 5, 7, etc., but must be modest in comparison to the image dimension.
- During the forward pass, we slide each filter across the whole input volume step by step (each step is called stride and can have a value of 2 or 3 or even 4 for high dimensional images) and compute the dot product between the weights of the filters and the patch from the input volume.

- As we slide our filters, we'll get a 2-D output for each filter, which we'll stack together to generate an output volume with a depth equal to the number of filters. All of the filters will be learned by the network.

**Types of layers:**

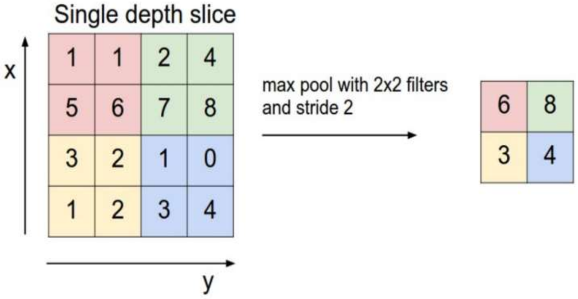
Let's take an example by running a covnets on of image of dimension 32 x 32 x 3.

1. Input Layer: This layer contains the raw image input with width 32, height 32, and depth 3.

2. Convolution Layer: The output volume is computed by computing the dot product of all filters and the image patch. If we utilize a total of 12 filters for this layer, the output volume will be 32 x 32 x 12.

3. Activation Function Layer: This layer will apply an element-by-element activation function to the convolution layer's output. RELU: max(0, x), Sigmoid: 1/(1+e-x), Tanh, Leaky RELU, and others are frequent activation functions. Because the volume remains constant, the output volume will be 32 x 32 x 12.

4. Pool Layer: This layer is added into the covnets on a regular basis, and its major job is to lower the size of the volume, which speeds up computation, saves memory, and prevents overfitting. Pooling layers are classified into two types: maximum pooling and average pooling. When we utilize a max pool with 2 x 2 filters and stride 2, the resulting volume is 16x16x12.



|   |   |   |   |
|---|---|---|---|
| 1 | 1 | 2 | 4 |
| 5 | 6 | 7 | 8 |
| 3 | 2 | 1 | 0 |
| 1 | 2 | 3 | 4 |

|   |   |
|---|---|
| 6 | 8 |
| 3 | 4 |

**Figure 4: Representation of max pooling**

5. Fully-Connected Layer: This layer is a typical neural network layer that takes input from the previous layer, computes class scores, and outputs a 1-D array with the number of classes as the size.

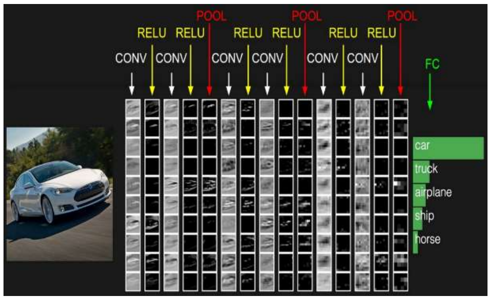
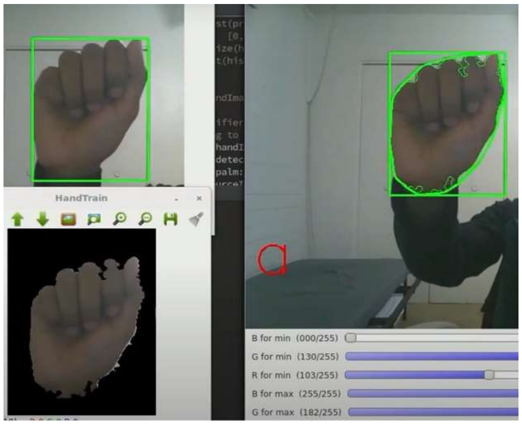


**Figure 5: Output in the form of 1-D array**

## V. RESULTS

Below figures explains the snapshots for Sign Language Recognition using CNN.



B for min (000/255)
G for min (130/255)
R for min (103/255)
B for max (255/255)
G for max (182/255)

**Figure 6: The alphabet 'a' depicted in ASL.**

Fig 6. This screenshot shows the result of sign language recognition, with the user displaying the letter 'a' in American Sign Language.

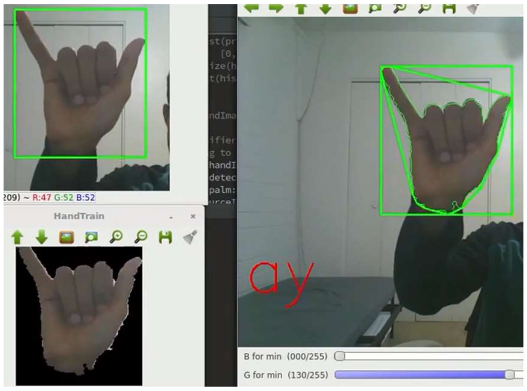


**Figure 7: The alphabet ‘y’ depicted in ASL.**

Fig 7. This screenshot shows the result of sign language recognition, in which the user presents the letter 'y' in American Sign Language, and the supplied letter is appended to the preceding letter to make "ay".

## VI. CONCLUSION

Convolution Neural Networks (CNNs) are being utilized in this work to recognize alphabets in sign language. Sign language is a visual language that communicates using gestures, facial expressions, and body language. The CNN method is used to recognize and categories sign language hand motions, which may subsequently be mapped to particular letters of the alphabet. Machine learning is an artificial intelligence area that includes training algorithms to make predictions or judgements based on data patterns. In this scenario, the CNN algorithm is being trained on a dataset of sign language motions and their accompanying alphabet letters to learn how to recognize and categorize new movements.

## REFRENCES

[1] Zhi-hua Chen, Jung-Tae Kim, Jianning Liang, Jing Zhang, Yu-Bo Yuan, "Real-Time Hand Gesture Recognition Using Finger Segmentation", *The Scientific World Journal*, vol. 2014, Article ID 267872, 9 pages, 2014. https://doi.org/10.1155/2014/267872

[2] A. Kumar and A. Kumar, “Dog Breed Classifier for Facial Recognition using Convolutional Neural Networks,” pp. 508–513, 2020.

[3] S. Y. Kim, H. G. Han, J. W. Kim, S. Lee, and T. W. Kim, “A hand gesture recognition sensor using Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021) IEEE Xplore Part Number

[4] Y. Cui and J. Weng, “A learning-based predictionand-verification segmentation scheme for hand sign image sequence,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 21, no. 8, pp. 798–804, 1999, doi: 10.1109/34.784311.

[5] K. L. Bouman, G. Abdollahian, M. Boutin, and E. J. Delp, “A low complexity sign detection and text localization method for mobile applications,” IEEE Trans. Multimed., vol. 13, no. 5, pp. 922–934, 2011, doi: 10.1109/TMM.2011.2154317.

[6] J. Wu, L. Sun, and R. Jafari, “A Wearable System for Recognizing American Sign Language in RealTime Using IMU and Surface EMG Sensors,” IEEE J. Biomed. Heal. Informatics, vol. 20, no. 5, pp. 1281–1290, 2016, doi: 10.1109/JBHI.2016.2598302.

[7] C. Zhang, W. Ding, G. Peng, F. Fu, and W. Wang, “Street View Text Recognition with Deep Learning for Urban Scene Understanding in Intelligent Transportation Systems,” IEEE Trans. Intell. Transp. Syst., pp. 1–17, 2020, doi: 10.1109/tits.2020.3017632.

[8] M. Safeel, T. Sukumar, K. S. Shashank, M. D. Arman, R. Shashidhar, and S. B. Puneeth, “Sign Language Recognition Techniques- A Review,” 2020 IEEE Int. Conf. Innov. Technol. INOCON 2020, pp. 1–9, 2020, doi: 10.1109/INOCON50539.2020.9298376.

[9] W. Aly, S. Aly, and S. Almotairi, “Userindependent american sign language alphabet recognition based on depth image and PCANet features,” IEEE Access, vol. 7, pp. 123138–123150, 2019, doi: 10.1109/ACCESS.2019.2938829.

[10] S. Adhikari, S. Thapa, and B. K. Shah, “Oversampling based Classifiers for Categorization of Radar Returns from the Ionosphere,” Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020, no. Icesc, pp. 975–978, 2020, doi: 10.1109/ICESC48915.2020.9155833.

[11] S. Hayani, M. Benaddy, O. El Meslouhi, and M. Kardouchi, “Arab Sign language Recognition with Convolutional Neural Networks,” Proc. 2019 Int. Conf. Comput. Sci. Renew. Energies, ICCSRE 2019, pp. 2019–2022, 2019, doi: 10.1109/ICCSRE.2019.8807586.

[12] Y. Li, X. Wang, W. Liu, and B. Feng, “Pose Anchor: A Single-Stage Hand Keypoint Detection Network,” IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 7, pp. 2104–2113, 2020, doi: 10.1109/TCSVT.2019.2912620.

[13] X. Shen and F. L. Chung, “Deep network embedding for graph representation learning in signed networks,” arXiv, vol. 50, no. 4, pp. 1556– 1568, 2019.

[14] Y. Zhu, M. Liao, M. Yang, and W. Liu, “Cascaded Segmentation-Detection Networks for Text-Based Traffic Sign Detection,” IEEE Trans. Intell. Transp. Syst., vol. 19, no. 1, pp. 209–219, 2018, doi: 10.1109/TITS.2017.2768827.

[15] B. K. Shah, V. Kedia, R. Raut, S. Ansari, and A. Shroff, “Evaluation and Comparative Study of Edge Detection Techniques,” vol. 22, no. 5, pp. 6–15, 2020, doi: 10.9790/0661-2205030615.

[16] B. K. Shah, V. Kedia and R. K. Jha, "Integrated Vendor-Managed Time Efficient Application to Production of Inventory Systems," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 275-280, doi: 10.1109/ICICT50816.2021.9358504.

[17] Dr. Thamaraiselvi1, Challa Sai Hemanth and J Hruday Vikas, “Sign Language Recognition Using CNN” vol. 09 no. 03, e-ISSN: 2395-0056, p-ISSN: 2395-0072, Tamil Nadu, India, 2022.