VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Sign Language Recognition

Final Project Report

Submitted to Prof. Saravanakumar K

Course: CSE1019- Technical Answers to Real World Problems

Slot: TD1+TD2

Team -9 :

JYOTHI K C	**19BDS0144**
PARDHA SARADHI METTA	**19BCE0526**
ROHAN KUMAR	**19BCB0038**
GAUTHAM SREEKUMAR	**19BCE0818**
RANJIT CHOWDARY	**19BCI0193**

1 ABSTRACT

Man is of social nature and hence, communication is a fundamental skill necessary for survival in society. As of 2021, the world has roughly around 5% population, that is 466 million people , who have hearing or speech disabilities and this is estimated to shoot up to 900 million in the next 30 years. People with disabilities often face numerous obstacles from lack of widespread specialised learning facilities, employment opportunities to minimal provision of communication interfaces. In this project, deep learning models were used for recognising hand gestures used in the American sign language. The vision-based sign language detection system was designed using a publicly available benchmarked dataset , American sign language (ASL) Alphabet dataset. Two convolutional neural network(CNN) models with architectural variations were used: ResNet-34 and EfficientNetB0. To overcome the computational overhead of the traditional CNN model, the CNN models with variant architectures ,the ResNet and EfficientNet, was implemented on the ASL Alphabet dataset. The ResNet and EfficientNet models achieved high accuracy scores of 0.9956 and 0.9995 respectively. The proposed approach for real-time sign language recognition , can be used to develop a translator or gesture-to-speech input tools and features in applications.

2 INTRODUCTION

Gesture, a symbol of physical behavior or emotional expression includes body gesture and hand gesture. Gesture recognition is the ability of a computer to understand human gestures and execute commands based on those gestures. In gesture recognition, real-time image data is fed by a camera into a sensing device that is connected to a computer. Using a predetermined gesture library, the software identifies meaningful gestures where each gesture is matched to a computer command. The software then correlates the real-time gesture, interprets the gesture and uses the library to identify meaningful gestures that match the library. Human Computer Interaction aims to improve the interaction between humans and the computers and nowadays it is not just limited to keyboard and mouse interaction (Haria et al., 2017). In hand gesture recognition, sensors are used to read and interpret hand movements as commands.

Hand gestures can be broadly classified as static and dynamic gestures. Static hand gestures are those gestures where the position and orientation of the hand in space does not change for a given time, while if there are changes either within the given time, they are called dynamic gestures. Joining the thumb and the forefinger to express the “Ok” symbol is a static gesture. Waving of the hand is an example of dynamic gesture. Vision based approaches for hand gesture recognition require no hand devices. But in a non-vision-based approach, some devices are involved in recognition. Use of a pair of wired gloves in the detection of finger movement is an example of this approach.

Hand gesture recognition is one of the most advanced domains in human computer interaction and it has several real world applications such as games, sign language recognition, assisted living, virtual reality, robot control, physical exercise monitoring etc (Panwarand Singh, 2011; Tan and Gup, 2011; Droeschelet al., 2011). Being a natural means of interaction, they are commonly used for communication purposes by speech impaired people worldwide. It can also be used for controlling other systems within a vehicle like heating and cooling, control of smart home systems and many more (Huu et al., 2009). By connecting with telematics systems, hand gesture identification can be used to allow the vehicle to provide information about the nearby landmarks if it recognizes that an occupant is pointing at it. Improved safety is one of the most important benefits of ge Gesture, a symbol of physical behavior or emotional expression includes body gesture and hand gesture. Gesture recognition is the ability of a computer to understand human gestures and execute commands based on those gestures. In gesture recognition, real-time image data is fed by a camera into a sensing device that is connected to a computer. Using a predetermined gesture library, the software identifies meaningful gestures where each gesture is matched to a computer command. The software then correlates the real-time gesture, interprets the gesture and uses the library to identify meaningful gestures that match the library. Human Computer Interaction aims to improve the interaction between humans and the computers and nowadays it is not just limited to keyboard and mouse interaction (Haria et al., 2017). In hand gesture recognition, sensors are used to read and interpret hand movements as commands.

Hand gestures can be broadly classified as static and dynamic gestures. Static hand gestures are those gestures where the position and orientation of the hand in space does not change for a given time, while if there are changes either within the given time, they are called dynamic gestures. Joining the thumb and the forefinger to express the “Ok” symbol is a static gesture. Waving of the hand is an example of dynamic gesture. Vision based approaches for hand gesture recognition require no hand devices. But in a non-vision-based approach, some devices are involved in recognition. Use of a pair of wired gloves in the detection of finger movement is an example of this approach.

Hand gesture recognition is one of the most advanced domains in human computer interaction and it has several real world applications such as games, sign language recognition, assisted living, virtual reality, robot control, physical exercise monitoring etc (Panwarand Singh, 2011; Tan and Gup, 2011; Droeschelet al., 2011). Being a natural means of interaction, they are commonly used for communication purposes by speech impaired people worldwide. It can also be used for controlling other systems within a vehicle like heating and cooling, control of smart home systems and many more (Huu et al., 2009). By connecting with telematics systems, hand gesture identification can be used to allow the vehicle to provide information about the nearby landmarks if it recognizes that an occupant is pointing at it. Improved safety is one of the most important benefits of gesture recognition.

Gesture, a symbol of physical behavior or emotional expression includes body gesture and hand gesture. Gesture recognition is the ability of a computer to understand human gestures and execute commands based on those gestures. In gesture recognition, real-time image data is fed by a camera into a sensing device that is connected to a computer. Using a predetermined gesture library, the software identifies meaningful gestures where each gesture is matched to a computer command. The software then correlates the real-time gesture, interprets the gesture, and uses the library to identify meaningful gestures that match the library. Human Computer Interaction aims to improve the interaction between humans and the computers and nowadays it is not just limited to keyboard and mouse interaction (Haria et al., 2017). In hand gesture recognition, sensors are used to read and interpret hand movements as commands.

Hand gestures can be broadly classified as static and dynamic gestures. Static hand gestures are those gestures where the position and orientation of the hand in space does not change for a given time, while if there are changes either within the given time, they are called dynamic gestures. Joining the thumb and the forefinger to express the "Ok" symbol is a static gesture. Waving of the hand is an example of dynamic gesture. Vision based approaches for hand gesture recognition require no hand devices. But in a non-vision-based approach, some devices are involved in recognition. Use of a pair of wired gloves in the detection of finger movement is an example of this approach.

Hand gesture recognition is one of the most advanced domains in human computer interaction and it has several real-world applications such as games, sign language recognition, assisted living, virtual reality, robot control, physical exercise monitoring etc (Panwarand Singh, 2011; Tan and Gup, 2011; Droeschelet al., 2011). Being a natural means of interaction, they are commonly used for communication purposes by speech impaired people worldwide. It can also be used for controlling other systems within a vehicle like heating and cooling, control of smart home systems and many more (Huu et al., 2009). By connecting with telematics systems, hand gesture identification can be used to allow the vehicle to provide information about the nearby landmarks if it recognizes that an occupant is pointing at it. Improved safety is one of the most important benefits of gesture recognition.

2.1 DEEP LEARNING MODELS USED:

2.1.1 CNN

Convolution neural network (CNN) is widely used in image detection and recognition because it can recognize features regardless of the appeared position. In a neural network, each node in the previous layer gives effects to all nodes in the next layer. However, in CNN, only several nodes in the current layer give effects to the nodes in the next layer. So, CNNs can use local correlation. It means that CNN learns features from the images. CNN, it has two special layers such as a convolution layer and a pooling layer. In the convolution layer, features are extracted by convoluting filter to inputs. In the

pooling layer, an input is down sampled to decrease the effect of small position shifting. CNN is consisted of some sets of these two types of special layers and normal NN. By using the pooling layer, the deep learning model is robust to the minor changes in the images. A critical feature of CNNs is their ability to achieve 'spatial invariance,' which implies that they can learn to recognize and extract image features anywhere in the image. There is no need for manual extraction as CNNs learn features by themselves from the image/data and perform extraction directly from images. This makes CNNs a potent tool within Deep Learning for getting accurate results.

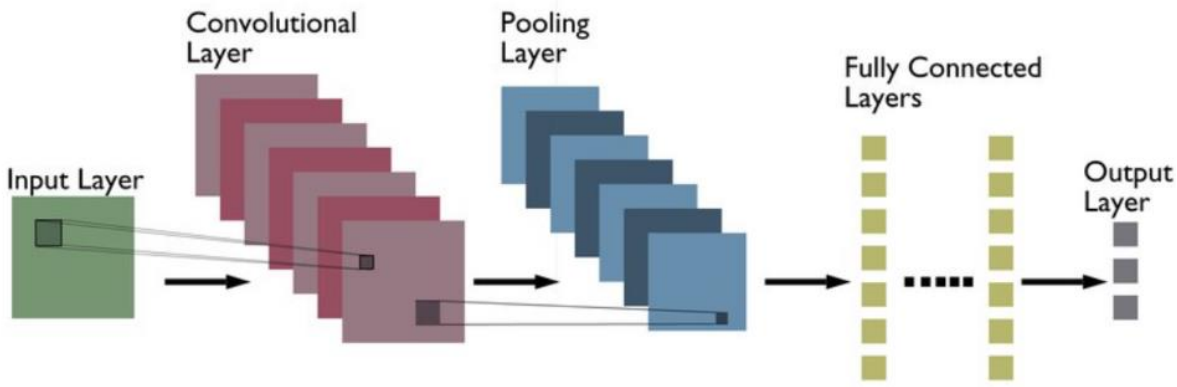

Figure 1

The following are definitions of different layers shown in the above architecture:

Convolutional layer: Convolutional layers are made up of a set of filters (also called kernels) that are applied to an input image. The output of the convolutional layer is a feature map, which is a representation of the input image with the filters applied. Convolutional layers can be stacked to create more complex models, which can learn more intricate features from images.

Pooling layer: Pooling layers are a type of convolutional layer used in deep learning. Pooling layers reduce the spatial size of the input, making it easier to process and requiring less memory. Pooling also helps to reduce the number of parameters and makes training faster. There are two main types of pooling: max pooling and average pooling. Max pooling takes the maximum value from each feature map, while average pooling takes the average value. Pooling layers are typically used after convolutional layers in order to reduce the size of the input before it is fed into a fully connected layer.

Fully connected layer: Fully-connected layers are one of the most basic types of layers in a convolutional neural network (CNN). As the name suggests, each neuron in a fully-connected layer is Fully connected- to every other neuron in the previous layer. Fully connected layers are typically used towards the end of a CNN- when the goal is to take the features learned by the previous layers and use

them to make predictions. For example, if we were using a CNN to classify images of animals, the final Fully connected layer might take the features learned by the previous layers and use them to classify an image as containing a dog, cat, bird, etc.

2.1.2 ResNet

Residual network(ResNet) is a CNN architecture that was developed by Kaiming He et al. It comprises of 32 layers and over one million parameters. The ResNet model is widely used to solve natural language processing problems like sentence completion or machine comprehension. A few real-life applications of ResNet CNN architecture include Microsoft's machine comprehension system, which has used CNNs to generate the answers for more than 100k questions in over 20 categories. The CNN architecture ResNet is computationally efficient and can be scaled up or down to match the computational power of GPUs. To solve the problem of the vanishing/exploding gradient, this architecture introduced the concept called Residual Blocks. The skip connection method connects activations of a layer to further layers by skipping some layers in between. This forms a residual block. Resnets are made by stacking these residual blocks together.

The approach behind this network is instead of layers learning the underlying mapping, we allow the network to fit the residual mapping. So, instead of say H(x), initial mapping, let the network fit,

$$F(x) := H(x) - x \; which \; gives \; H(x) := F(x) + x$$

The algorithm can be implemented using the Python libraries Tensorflow and Keras API. Figure 2 represents the architecture of a ResNet-34 model.

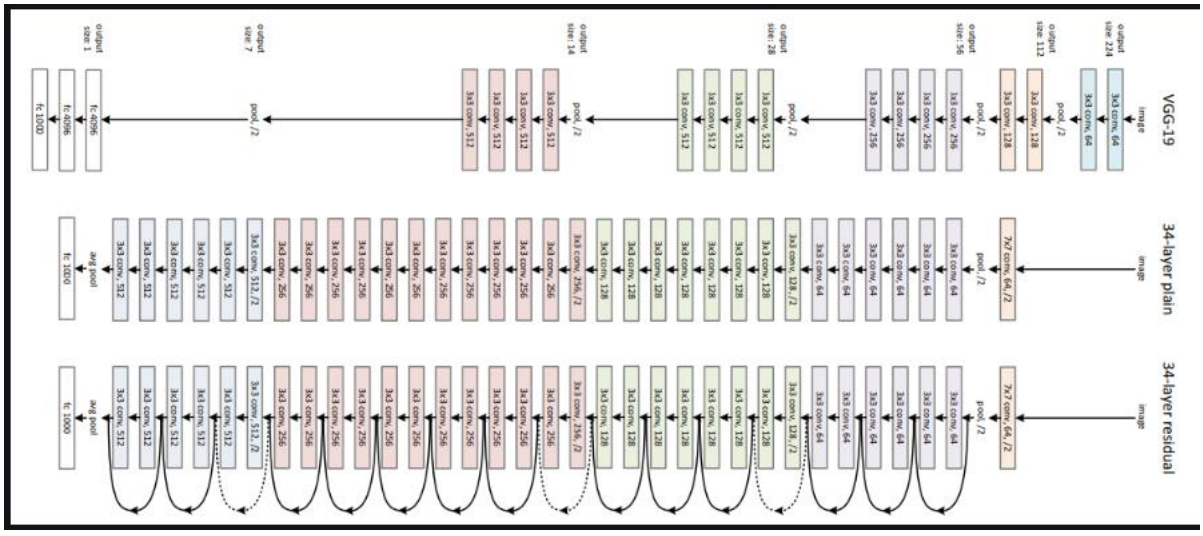

Figure 2 Architecture of ResNet-34 model

2.1.3 EfficientNet

EfficientNet is a convolutional neural network architecture and scaling method that uniformly scales all dimensions of depth/width/resolution using a compound coefficient. Unlike conventional practice

that arbitrary scales these factors, the EfficientNet scaling method uniformly scales network width, depth, and resolution with a set of fixed scaling coefficients. The compound scaling method is justified by the intuition that if the input image is bigger, then the network needs more layers to increase the receptive field and more channels to capture more fine-grained patterns on the bigger image. The base EfficientNet-B0 network is based on the inverted bottleneck residual blocks of MobileNetV2, in addition to squeeze-and-excitation blocks. Figure 3 depicts an architecture diagram of the model.

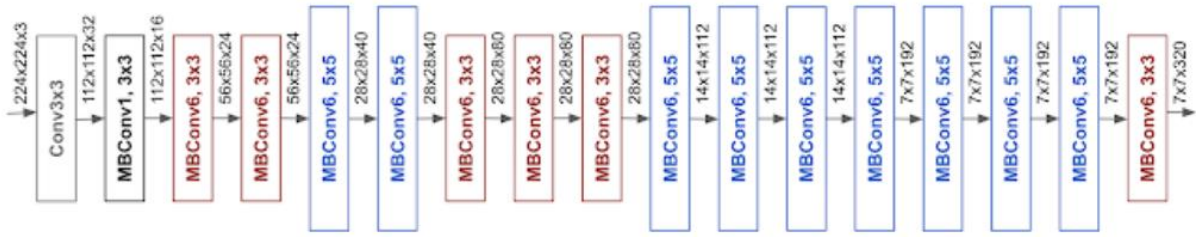

Figure 3 architecture of EfficientNet-B0

3 RELATED LITERATURE

Using a set of features, a 3-D representation of the angular relationship between a dozen biological components was created (Illuri et al., 2022). Hidden Markov models were then applied to a feature vector in order to encode it. An in-depth explanation of the process of developing a transition gesture model was presented in order to accurately recognize significant movements. The findings demonstrate that hand postures may be recognized with a minimum of inaccuracy when using a computer. According to the experimental results, the system successfully recognized hand motions, and the system's performance was suitable for real-time implementations. The best performing models CNN achieved higher accuracy and lower loss for training and validation datasets when compared to the baseline performance model. The validation loss is less than the training data loss; however, the accuracy of the validation dataset is more or less higher than the training dataset, which indicates some discrepancy in data.

The advent of deep learning has revolutionized the domain of computer vision. Convolutional neural networks (CNNs) became state-of-the-art for solving complex tasks thanks to technological advances of high-end accelerators, such as GPUs and FPGAs, combined in clusters or cloud solutions. In embedded systems, CNNs are also of great interest. Heidorn et al., (2021) designed a CNN-based hand sign recognition which is accelerated on a TCPA,implemented the TCPA prototypically as an overlay on a Xilinx Zynq System-on-a-Chip (SoC), and showcase tremendous speed ups compared with the integrated ARM Cortex-A9 processor.

Sign Language is used by the deaf and voiceless community to be able to communicate with others, but the most commonly faced problem here is that everyone around may not be able to understand sign language. Pala et al., (2021) designed a system with a motive to bridge the communication gap between the communities, therefore, establish the interaction between the speechless community to communicate with others. Hand gestures differ from one person to another person in shape and orientation; therefore, a problem of linearity arose. In this paper, a comparison between KNN, SVM, and CNN algorithms has been done to determine which algorithm would provide the best accuracy among all. Approximately 29,000 images were split into test and train data and pre-processed to fit into the KNN, SVM, and CNN models to obtain an accuracy of 93.83%, 88.89%, and 98.49% respectively. However, the training time of models was long due to the large dataset. Some big data processing methods can be implemented to rectify this issue.

Cui and Weng (1996) presented a prediction-and-verification segmentation scheme wing attention imagesfrom multiple fixations. A major advantage of this scheme was that it could handle a large number of different deformable objects presented in complex backgrounds. The scheme was also relatively efficient since the segmentation was guided by the past knowledge through a prediction-and-verification scheme. The system was tested to segment hands in the sequences of intensity images, where each sequence represented a hand sign. The experimental result showed a 95% correct segmentation rate with a 3% false rejection rate.

Sign and gesture recognition offers a natural way for human–computer interaction. In the work of Ibarguren et al. (2010), a real time sign recognition architecture including both gesture and movement recognition has been presented. Among the different technologies available for sign recognition data gloves and accelerometers were chosen for the purposes of this research. Due to the real time nature of the problem, the proposed approach works in two different tiers, the segmentation tier and the classification tier. In an effort to emphasize the real use of the architecture, this approach deals specially with problems like sensor noise and simplification of the training phase. However, in this model, when there was a slight hand movement between two signs, the sign performed in two parts, or the sign was corrected, segmenting problems still existed. Creation of more complex classifiers seems to be one the best choices to improve recognition rates.

Hand sign language recognition from video is a challenging research area in computer vision, which performance is affected by hand occlusion, fast hand movement, illumination changes, or background complexity, just to mention a few. In recent years, deep learning approaches have achieved state-of-the-art results in the field, though previous challenges are not completely solved. Rastgoo et al., (2020) proposed a deep learning-based pipeline architecture for efficient automatic hand sign language recognition using Single Shot Detector (SSD), 2D Convolutional Neural Network (2DCNN), 3D Convolutional Neural Network (3DCNN), and Long Short-Term Memory (LSTM) from RGB input

videos. They have used a CNN-based model which estimates the 3D hand keypoints from 2D input frames. The model showed improved state of the art results for hand sign language.

She et al., (2014) presented a literature review on hand tracking and gesture recognition. The survey examined 37 papers describing depth-based gesture recognition systems in terms of (1) the hand localization and gesture classification methods developed and used, (2) the applications where gesture recognition has been tested, and (3) the effects of the low-cost Kinect and OpenNI software libraries on gesture recognition research. The papers that use the Kinect and the OpenNI libraries for hand tracking tend to focus more on applications than on localization and classification methods and showed that the OpenNI hand tracking method is good enough for the applications tested thus far. However, the limitations of the Kinect and other depth sensors for gesture recognition have yet to be tested in challenging applications and environments.

Tan and Guo (2011) introduced a hand gesture recognition system to recognize the alphabets of Indian Sign Language. In the proposed system, there were 4 modules: real time hand tracking, hand segmentation, feature extraction and gesture recognition. Camshift method and Hue, Saturation, Intensity (HSV) color model were used for hand tracking and segmentation. For gesture recognition, genetic algorithm was used. They proposed an easy-to-use and inexpensive approach to recognize single handed as well as double handed gestures accurately.

Communications between deaf-mute and a normal person have always been a challenging task. A glove based deaf-mute communication interpreter system was proposed by Bhujbal and Warhade (2018) to facilitate such people. The glove was internally equipped with five flex sensors, tactile sensors and accelerometer. For each specific gesture, the flex sensor produces a proportional change in resistance and accelerometer measures the orientation of hand. The processing of these hand gestures was done in Arduino. The glove included two modes of operation – training mode to benefit everyuser and an operational mode. The concatenation of letters to form words was also done in Arduino. In addition, the system also included a text to speech conversion (TTS) block which translates the matched gestures i.e., text to voice output.

According to the census of India 2011, 70 million people have some kind of disability, among of that 18% of people are speech and hear impaired. This means that India is the country which has a large number of people having this kind of disability. These people experience the problem to participate in society and the enjoyment of equal rights and opportunities because they don't have the power to express feelings in the form of words and sentence. In a system proposed by Taniya et al. (2019), processor collects data from the 5 flex sensors and accelerometer. Further, processor matches the data which is received from the sensors and the previously saved data. If data matched with the saved data then assigned meaning for that data will be displayed on LCD screen and also send it to the Android mobile through Bluetooth. Android mobile app can convert this into voice. So, this system has

an ability to convert sign language into a voice in a very simple way. Further steps can be taken to tackle the computation and complexity issues of deep learning models such as MLP.

Among many of the fastest growing research fields, sign language recognition is one of the top. Sign Language is a methodical coded language where meanings are assigned to every gestures. To create a strong interface between user and computer, recognition of gesture is important. In a study by Chen et al. (2014), a hand gesture recognition method based on multiscale density features is proposed. Depth images of numerals of American Sign Language were considered in this work and recognition rate of 98.20% was obtained, which was comparable with related state-of- the art methods.

The intention of study by Katoch et al., (2022) was to discuss hand gesture recognition based on detection of some shape based features. The set up consisted of a single camera to capture the gesture formed by the user and take this hand image as an input to the proposed algorithm. The overall algorithm was divided into four main steps, which includes segmentation, orientation detection, feature extraction and classification. The proposed algorithm is independent of user characteristics. It does not require any kind of training of sample data. The proposed algorithm was tested on 390 images, with a recognition rate of approximately 92% and average elapsed time of 2.76 sec. It takes a less computation time as compared to other approaches.

Using gestures can help people with certain disabilities in communicating with other people. Haria et al., (2017) proposed a lightweight model based on YOLO (You Only Look Once) v3 and DarkNet-53 convolutional neural networks for gesture recognition without additional preprocessing, image filtering, and enhancement of images. They achieved better results by extracting features from the hand and recognized hand gestures of the proposed YOLOv3 based model with accuracy, precision, recall, and an F-1 score of 97.68, 94.88, 98.66, and 96.70%, respectively. Further, they compared the model with Single Shot Detector (SSD) and Visual Geometry Group (VGG16), which achieved accuracy between 82 and 85%. The trained model can be used for real-time detection, both for static hand images and dynamic gestures recorded on a video.

Miller et al., (2020) showed how surgeons can interact with medical images using finger and handgestures in two situations: one hand-free and no hands-free interaction. The system permits the following important capabilities: (1) touch-less input for sterile interaction with connected health applications, (2) hand and finger gesture recognition when either one or both hands are busy holding tools, extending multitasking capabilities for health professionals, and (3) mobile and networked, allowing for custom wearable and non-wearable configurations. They evaluated the system in a simulated operating room to manipulate preoperative images using four gestures: circle, double tap, swipe, and finger click. They collected data from five subjects and trained a K- Nearest-Neighbor multi-class classifier using 15-fold cross validation, achieving a 94.5% precisionfor gesture classification and

concluded that the proposed system performed with high accuracy and is useful in cases where only one hand or a few fingers are free to interact when the hands are busy.

Katoch et al., (2022) recently presented a technique that uses the Bag of Visual Words model (BOVW) to recognize Indian sign language alphabets (A-Z) and digits (0–9) in a live video stream and output the predicted labels in the form of text as well as speech. Segmentation is done based on skin colour as well as background subtraction. SURF (Speeded UpRobust Features) features were extracted from the images and histograms were generated to map the signs with corresponding labels. The Support Vector Machine (SVM) and Convolutional Neural Networks (CNN) were used for classification. An interactive Graphical User Interface (GUI) was also developed for easy access corresponding labels.

4 PROBLEM STATEMENT

Hand gesture recognition is one of the most advanced domains in human computer interaction and it has several real world applications. Several approaches have been reported for hand gesture recognition with variable accuracy levels. The proposed models have several drawbacks such as less accuracy, longer training time for large datasets, image segmenting problems, lack of complex classifiers and so on. In this paper, we aim to implement a vision based, static hand gesture recognition model using machine learning. Vision based approaches for hand gesture recognition require no hand devices. But in a non-vision-based approach, some devices are involved in recognition. A critical feature of CNNs is their ability to achieve 'spatial invariance', which implies that they can learn to recognize and extract image features anywhere in the image. There is no need for manual extraction as CNNs learn features by themselves from the image/data and perform extraction directly from images. This makes CNNs a potent tool within Deep Learning for getting accurate results. Another advantage of CNNs is their ability to develop an internal representation of a two-dimensional image. This allows the model to learn position and scale in variant structures in the data which is important when working with images. So, in this study, we propose CNN model with different architectures for solving the Sign Language (or Gesture) and their comparison.

5 PROPOSED METHODOLOGY

5.1 DATASETS

American Sign Language(ASL)

The image data set for alphabets in the American Sign Language was used in the study to predict hand sign recognition (https://github.com/grassknoted/Unvoiced). The dataset is a collection of images of alphabets from the American Sign Language, separated in 29 folders which represent the various

classes. The training data set contains 87,000 images which are 200×200 pixels. There are 29 classes, of which 26 are for the letters A-Z and 3 classes for *SPACE*, *DELETE* and *NOTHING*.

Hand gesture recognition image dataset(Hagrid)

HaGRID (HAnd Gesture Recognition Image Dataset) is one of the largest datasets for HGR systems. This dataset contains 552,992 full HD RGB images divided into 18 classes of functional gestures.

5.2 MODULE DESCRIPTION

Traditional CNN, Resnet34 and Efficientnet-b0 were the pre-trained architectures used for hand sign recognition. The code was written in python and pytorch library was used for training the models. The optimizer used was Adam optimizer with a Learning rate of 1e-3. Cosine annealed warm restart learning scheduler was employed in the study. The Loss function used was cross entropy loss and Evaluation matrix was Accuracy, for both the models.

5.3 ARCHITECTURE

Figure 4 represents the hierarchical task analysis of our proposed methodology.

Input The Input device used for the experiment was the camera which captures 2D images from the external world and passes it via an internal stream to the main algorithm which handles processing.

Hand shape mapping The frame data is extracted from the input stream. The image is then binarized based on the skin tone to distinguish the hand shape. The finger joints are then identified and mapped together.

Gesture Recognition The classifier models, CNN, is loaded and map the identified finger joint positions to the images fed during training. Currently the program follows a single user architecture that allows only for recognition and transcription of gesture data from one user. The ASL Alphabet train: test ratio is used as input for the training and testing of the ResNet and EfficientNet-B0 models.

Output All information was displayed directly on the user's computer screen, so the only output device used was a screen to display a 2D image of the camera feed with the gesture details overlaid on top of it.

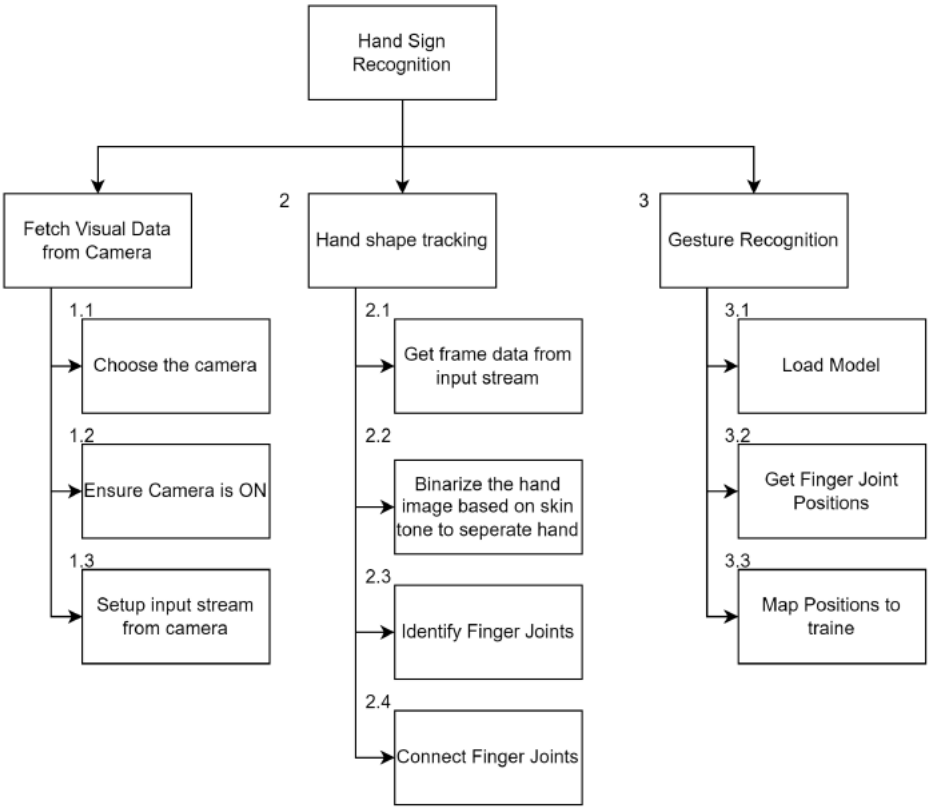

Figure 4

The flow of data in the proposed architecture is represented in three levels as shown in Figures 5.1,5.2 and 5.3.

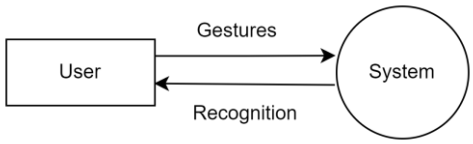

Figure 5.1. Level – 0

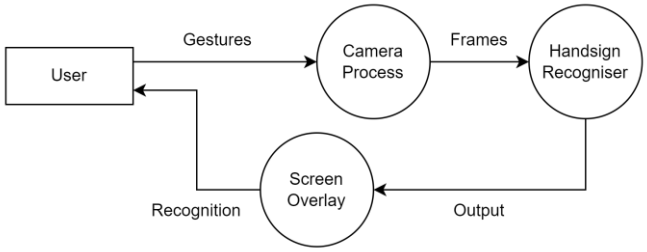

Figure 5.2. Level – 1

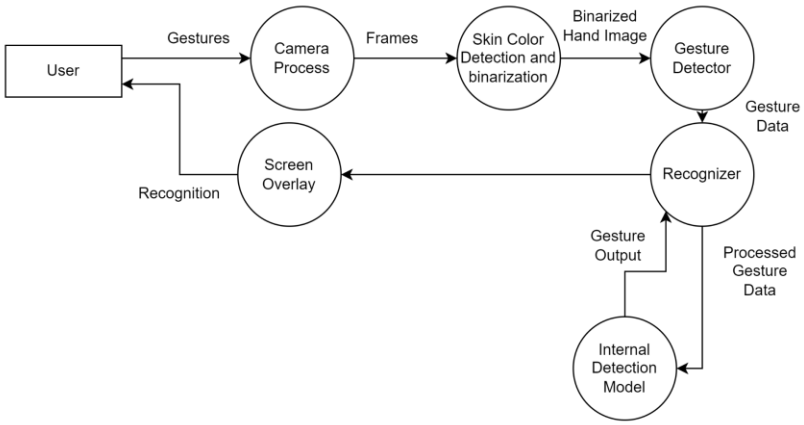

Figure 5.3 Level-3

Figure 5. Data Flow Diagram

5.4 ALGORITHM

The below steps were followed during the architectural implementation.

Pseudocode:

Initialize camera feed

Load models into memory

Load classes from models

Run an infinite while loop:

Check if (q) input provided to exit

If (q):

break

Read frame

Flip the frame

Change colour code to RGB if not already

Process the frame under hand recognition to separate the hand from the background

Run result through point and line recogniser to break down the hand gesture

For each point:

Draw point on overlay

For each line:

Draw line on overlay

Predict gesture based on points and lines

Convert prediction ID to corresponding name

Output camera frame to screen

Output overlay to screen

Close all windows

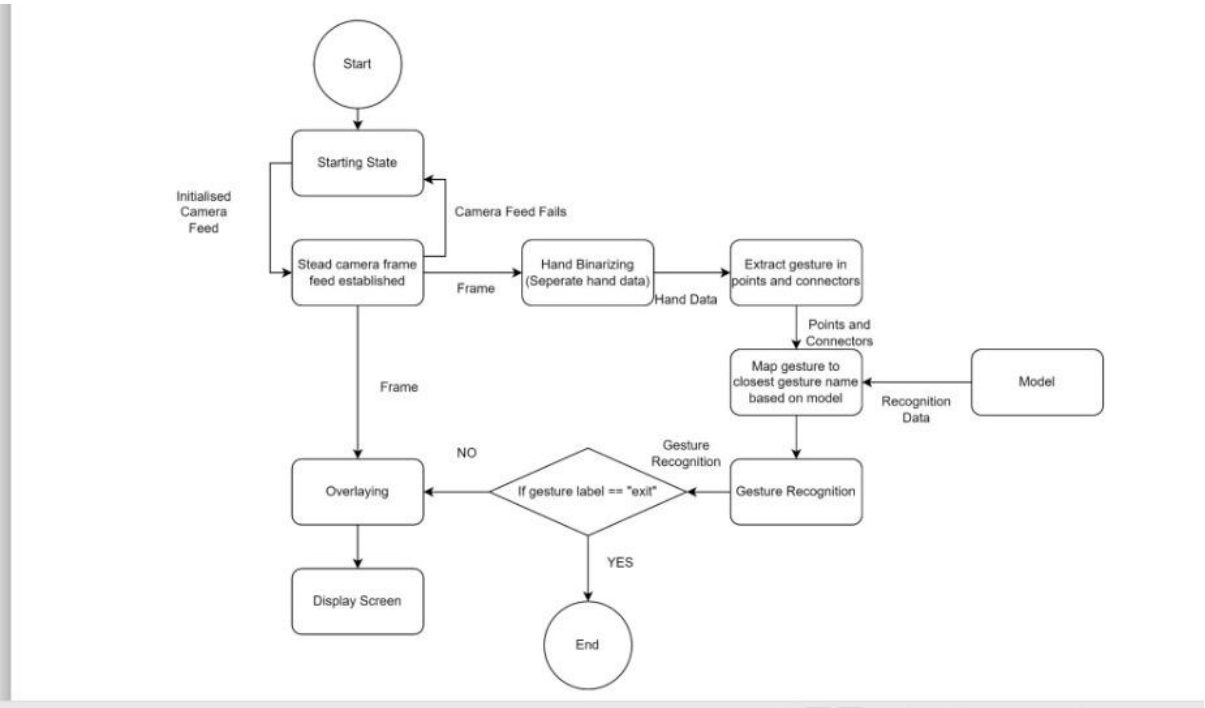

Figure 5 Flowchart of proposed methodology

6 RESULTS

The performance metric used for evaluation of the deep learning models is Accuracy. The two models, viz., Resnet34 and Efficientnet-b0 were employed on the ASL dataset to predict the hand signs with an accuracy of 99.96% and 99.995%, respectively. The Efficientnet-b0 was found to give more accuracy than the Resnet model. Figure 6 shows the accuracy of predictions using the two CNN models.

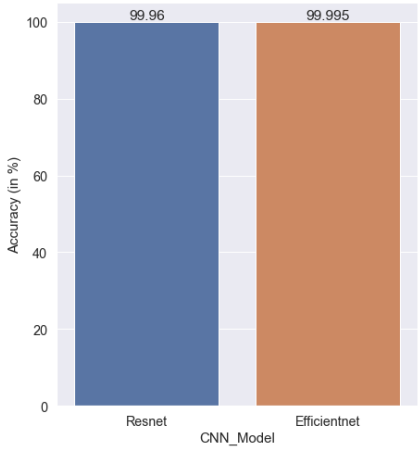

Figure 7. Accuracy of the ResNet and EfficientNet models

The traditional CNN algorithm was implemented on HaGRID for real-time gesture recognition. The model was able to successfully identify and correcty classify the gestures. Figure 9 depicts the output results of the model.

Figure 7. Output screenshots of the EfficientNet-B0 and ResNet models

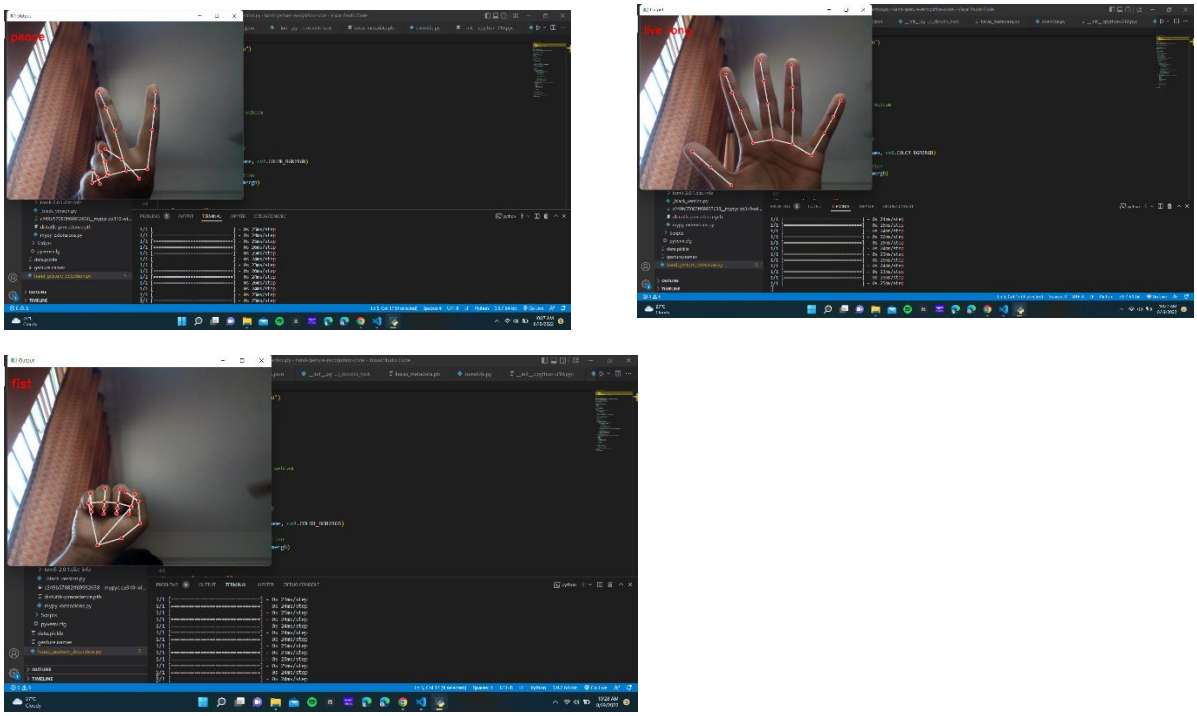

Figure 8. Output screenshots of the real-time traditional CNN model

7 CONCLUSION

The study showed that machine learning models are highly useful in hand sign recognition predictions. Both the models used here gave high accuracy of prediction, with Efficientnet-b0 giving comparatively more accuracy of 99.995%. Since these models doesn't use any sophisticated and costly input or output devices, they can be very effectively and economically used for gesture recognition, which has a lot of potential applications in an array of fields including sign language

recognition, smart home systems, vehicle control and many more. Adding more algorithms for comparison can also be scope for future work.

8 REFERENCES

1. What is gesture recognition? Gesture recognition defined | Marxent (marxentlabs.com).
2. Panwar, M. and Singh Mehra, P. "Hand gesture recognition for human computer interaction," *2011 International Conference on Image Information Processing*, 2011, pp. 1-7, doi: 10.1109/ICIIP.2011.6108940.
3. T.-D. Tan and Z.-M. Guo, “Research of hand positioning and gesture recognition based on binocular vision,” in *Proceedings of the IEEE International Symposium on Virtual Reality Innovations (ISVRI '11)*, pp. 311–315, March 2011.
4. D. Droeschel, J. Stückler, and S. Behnke, “Learning to interpret pointing gestures with a time-of-flight camera,” in *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI '11)*, pp. 481–488, March 2011.
5. Aashni Haria, Archanasri Subramanian, Nivedhitha Asokkumar, Shristi Poddar, Jyothi S Nayak. Hand Gesture Recognition for Human Computer Interaction, Procedia Computer Science, Volume 115, 2017, Pages 367-374, https://doi.org/10.1016/j.procs.2017.09.092.
6. https://www.aptiv.com/en/insights/article/what-is-gesture-recognition.
7. R. Yang, S. Sarkar, and B. Loeding, “Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 462–477, 2010.
8. Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, “American sign language recognition with the kinect,” in *Proceedings of the 13th ACM International Conference on Multimodal Interfaces (ICMI '11)*, pp. 279–286, November 2011.
9. A. Shimada, T. Yamashita, and R.-I. Taniguchi, “Hand gesture based TV control system—towards both user—& machine-friendly gesture applications,” in *Proceedings of the 19th Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV '13)*, pp. 121–126, February 2013.

10. C. Keskin, F. Kiraç, Y. E. Kara, and L. Akarun, “Real time hand pose estimation using depth sensors,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV '11)*, pp. 1228–1234, November 2011.

11. J. Zeng, Y. Sun, and F. Wang, “A natural hand gesture system for intelligent human-computer interaction and medical assistance,” in *Proceedings of the 3rd Global Congress on Intelligent Systems (GCIS '12)*, pp. 382–385, November 2012.

12. 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, ArticleID 267872, 9 pages, 2014. https://doi.org/10.1155/2014/267872.

13. Granit Luzhnica, Elizabeth Lex, Viktoria Pammer. A Sliding Window Approach to Natural Hand Gesture Recognition using a Custom Data Glove. In: 3D User Interfaces (3DUI); 2016 IEEE Symposium on 2016 Mar 19 ; New York : IEEE; 2016 ; p.81-90.

14. Hung CH, Bai YW, Wu HY. Home outlet and LED array lamp controlled by a smartphone with a hand gesture recognition. In: Consumer Electronics (ICCE); 2016 IEEE International Conference on ; 2016 Jan 7; New York : IEEE;2016 ; p.5-6.

15. Hung CH, Bai YW, Wu HY. Home appliance control by a hand gesture recognition belt in LED array lamp case. In: Consumer Electronics (GCCE); 2015 IEEE 4th Global Conference on ; 2015 Oct 27; New York : IEEE; 2015; p. 599-600.

16. She Y, Wang Q, Jia Y, Gu T, He Q, Yang B. A real-time hand gesture recognition approach based on motion features of feature points. In: Computational Science and Engineering (CSE); 2014 IEEE 17th International Conference on; 2014 Dec 19; New York: IEEE;2014;p.1096-1102.

17. Lee DH, Hong KS. A Hand gesture recognition system based on difference image entropy. In: Advanced Information Management and Service (IMS), 2010 6th International Conference on; 2010 Nov 30; Seoul; New York: IEEE; 2010 ; p. 410-413.

18. Hussain I, Talukdar AK, Sarma KK. Hand gesture recognition system with real-time palm tracking. In: India Conference (INDICON);2014 Annual IEEE ;2014 Dec 11; India, Pune; New York: IEEE; 2014; p. 1-6.

19. Huong TN, Huu TV, Le Xuan T. Static hand gesture recognition for vietnamese sign language (VSL) using principle components analysis. In: Communications, Management and

Telecommunications (ComManTel); 2015 International Conference on; 2015 Dec 28; p. 138-141.

20. Chen Y, Ding Z, Chen YL, Wu X. Rapid recognition of dynamic hand gestures using leap motion. In: Information and Automation; 2015 IEEE International Conference on; 2015 Aug 8; New York : IEEE;2015 ;p. 1419-1424.

21. Leem, S.K.; Khan, F.; Cho, S.H. Detecting Mid-air Gestures for Digit Writing with Radio Sensors and a CNN. IEEE Trans. Instrum. Meas. 2019, 69, 1066–1081.

22. Miller, E.; Li, Z.; Mentis, H.; Park, A.; Zhu, T.; Banerjee, N. RadSense: Enabling one hand and no hands interaction for sterile manipulation of medical images using Doppler radar. Smart Health 2020, 15, 100089.

23. Huu, P.N.; Minh, Q.T.; The, H.L. An ANN-based gesture recognition algorithm for smart-home applications. KSII Trans. Internet Inf. Syst. 2020, 14, 1967–1983.

24. Elmezain, M.; Al-Hamadi, A.; Appenrodt, J.; Michaelis, B. A hidden markov model-based isolated and meaningful hand gesture recognition. Int. J. Electr. Comput. Syst. Eng. 2009, 3, 156–163.

25. Nyirarugira, C.; Choi, H.-R.; Kim, J.; Hayes, M.; Kim, T. Modified levenshtein distance for real-time gesture recognition. In Proceedings of the 6th International Congress on Image and Signal Processing (CISP), Hangzhou, China, 16–18 December 2013.

26. Saqib, S.; Ditta, A.; Khan, M.A.; Kazmi, S.A.R.; Alquhayz, H. Intelligent dynamic gesture recognition using CNN empowered by edit distance. Comput. Mater. Contin. 2020, 66, 2061–2076. [CrossRef] 35. Al-Hammadi, M.; Muhammad, G.; Abdul, W.; Alsulaiman, M.; Bencherif, M.A.; Mekhtiche, M.A. Hand gesture recognition for sign language using 3DCNN. IEEE Access 2020, 8, 79491–79509.

27. B. Illuri, V. B. Sadu, E. Sathish, M. Valavala, T. L. D. Roy and G. Srilakshmi, "A Humanoid Robot for Hand-Sign Recognition in Human-Robot Interaction (HRI)," 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies(ICAECT), 2022, pp. 1-5, doi: 10.1109/ICAECT54875.2022.9808034.

28. C. Heidorn, D. Walter, Y. E. Candir, F. Hannig and J. Teich, "Hand Sign Recognition via Deep Learning on Tightly Coupled Processor Arrays," 2021 31st International Conference onField-

Programmable Logic and Applications (FPL), 2021, pp. 388-388, doi: 10.1109/FPL53798.2021.00079.

29. G. Pala, J. B. Jethwani, S. S. Kumbhar and S. D. Patil, "Machine Learning-based Hand Sign Recognition," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 356-363, doi: 10.1109/ICAIS50930.2021.9396030.

30. Y. Cui and J. J. Weng, "Hand segmentation using learning-based prediction and verification forhand sign recognition," Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1996, pp. 88-93, doi: 10.1109/CVPR.1996.517058.

31. A. Ibarguren, I. Maurtua, B. Sierra, Layered architecture for real time sign recognition: Hand gesture and movement, Engineering Applications of Artificial Intelligence, Volume 23, Issue 7,2010, Pages 1216-1228, ISSN 0952-1976, doi.org/10.1016/j.engappai.2010.06.001.

32. Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, Hand sign language recognition using multi-view hand skeleton, Expert Systems with Applications, Volume 150,2020,113336,ISSN0957 4174, doi.org/10.1016/j.eswa.2020.113336.

33. V. P. Bhujbal and K. K. Warhade, "Hand Sign Recognition Based Communication System for Speech Disable People," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, pp. 348-352, doi: 10.1109/ICCONS.2018.8663054.

34. Taniya, S., Soumi, P., Subhadip, B., &Ayatullah, M. (2019). Hand sign recognition from depth images with multi-scale density features for deaf mute persons. In International Conference onComputational Intelligence and Data Science.

35. Katoch, S., Singh, V., & Tiwary, U. S. (2022). Indian Sign Language recognition system using SURF with SVM and CNN. Array, 14, 100141.

36. arXiv:2201.10060 [cs.CV] (or arXiv:2201.10060v1 [cs.CV], https://doi.org/10.48550/arXiv.2 201.10060.

37. Kumar. (2022). “Different types of CNN Architectures Explained: Examples”. 12.04.2022. Vitalflux. https://vitalflux.com/different-types-of-cnn-architectures-explained-examples/

38. Pawan et al, (2022). “Residual Networks (ResNet) – Deep Learning” .15.06.2022. GeekforGeeks https://www.geeksforgeeks.org/residual-networks-resnet-deep-learning/