Bangla Hand Gesture Recognition Using Neural Networks

Muhammad Shahriar Zaman, Shekh Tanjil Sharif, Badrul Haider Chowdhury, Rabab Khan Rongon   
*Department of Computer Science and Engineering,*

*Faculty of Electrical & Computer Engineering,*

*Military Institute of Science and Technology,*

*Mirpur Cantonment, Dhaka-1216, Bangladesh*

Email- shahriarzaman445@gmail.com, sk.tanjilsrf@gmail.com, sohel1779@gmail.com, rababrongon@gmail.com

***Abstract—* Congenital deafness is the condition where an individual has to live without the ability of hearing. This unfortunate condition haunts the lives of millions of people worldwide since it is a challenge for them to run their daily lives with such a communication impediment. Sign languages are helpful in this regard because they use hand gestures to convey messages. Bangla Sign Language system is one of them with its own gesture set. But since most regular individuals cannot decipher such gestures an automated system has been designed which can identify these hand gestures into Bangla alphabets and eventually words.**

***Keywords—* Deaf, Mute, Bangla Sign Language, Fingerspelling, Neural Networks, BDSL49, VGG, MobileNet**

# **Introduction**

Around 14 million of Bangladeshis face congenital, complete or partial deafness. This ailment can affect children born to parents with no hearing impediments. Common causes of congenital deafness include preeclampsia, anoxia, birth injuries, jaundice and other Rh factor issues, maternal diabetes, maternal drug/alcohol abuse and high blood pressure [1][2]. Sign languages are there to alleviate this issue like the Bangla Sign Language finalized around the year 1994. But since most people aren’t aware of sign language practice specially Bangla sign language, the deaf-mute individuals who know this practice suffer a communication gap.

From the state-of-the-art we can deduce that machine learning algorithms especially neural network models have proved their proficiency in classifying hand gesture images. We look forward to creating basic convolutional neural network models as well as importing pretrained advanced models for classifying hand gesture images from the BDSL49 dataset.

Among the paper contents the introduction portion discusses the motivation and basic outline of our research work. After this the Literature Review section discusses the relevant literature of gesture recognition technologies with their achievements and shortcomings. The methodology portion after them explains the entire workflow of the project from start to end. The results and discussions section will demonstrate the results of our experiment and discuss various relevant findings procured during the experiment. The portion after this will talk about the limitations and future endeavors of our work. Ultimately, the conclusion portion will conclude the paper by discussing the degree of success we reached through this research work.

# **Literature Review**

In today's rapidly advancing world of information technology and communication, new systems are continuously being developed to enhance user experiences and replicate real-world scenarios. Gesture recognition is one of the areas undergoing significant improvement. With the rise of affordable depth sensors like Kinect and Intel RealSense, a considerable effort has been dedicated to creating effective algorithms for gesture recognition. Point Cloud technology represents a major step forward, but before effectively using point clouds, it is essential to assess their quality for optimal recognition performance.

As technology progresses, new techniques for gesture recognition are constantly being explored and refined. Generally, gestures can be classified into three main categories:

**Static 2D**: Simple shapes like a fist or extended fingers, ideal for basic gesture identification, but become complex when tracking is required. [3]

**Dynamic 2D**: An extension of Static 2D where hand trajectories and features are combined for recognition. [4]

**3D**: Revolutionized by depth sensors like Kinect, which capture depth in images, allowing pixel distance from surfaces to be measured, enhancing recognition alongside color images.

While hand gesture detection using hand joint skeleton data has become more prevalent in computer vision, it still presents significant challenges. Traditional approaches, like machine learning and conventional feature extraction methods, have mainly focused on designing effective feature descriptors. However, these techniques often struggle with the complexity of accurately recognizing and distinguishing gestures, especially in dynamic and unstructured environments. As a result, there is a growing need for more advanced models that can better interpret and utilize the intricate information provided by hand skeleton data. [5] [6][7]

Ohn-Bar et al. introduced a feature generation method for skeleton datasets by employing the Histogram of Oriented Gradients (HOG) algorithm along with a descriptor. They applied a linear Support Vector Machine (SVM) for classification after converting the features into a 2D array using HOG. Other researchers have proposed various feature extraction techniques, including the use of covariance matrices for joint locations, joint angles, and the exploration of 3D geometric relationships between joints, as well as examining intraclass variance to improve gesture recognition. [8, 9]

Ma et al. used an Unscented Kalman Filter (UKF) to minimize noise in hand skeleton data, combined with an LSTM for gesture classification, achieving 85.92% and 80.44% accuracy for 14 and 28 gestures in the DHG dataset, respectively. [10] Nunez et al. employed a hybrid CNN-LSTM model for recognizing temporal 3D poses, achieving notable accuracy levels. Chen et al. introduced MFA-Net, which integrates motion features and three RNNs, resulting in improved performance across the DHG and SHREC'17 datasets. Boulahia et al. presented handwriting-inspired features (Hif3D) for 3D skeleton-based gesture classification, achieving strong results on the DHG dataset. Recently, research has shifted towards using self-attention mechanisms to enhance gesture recognition, addressing long-range dependencies, an approach first applied by Vaswani et al. in natural language processing. [12,13]

The researchers utilized attention mechanisms at multiple levels, applying CNNs for individual time steps to enhance hand gesture recognition performance. This approach led to 89.20% accuracy for 14 gestures and 85.00% for 28 gestures on the DHG dataset. Testing on the SHREC’17 dataset produced even better results, achieving 93.60% accuracy for 14 gestures and 90.70% for 28 gestures. These outcomes demonstrate the effectiveness of attention-based techniques in sequential tasks like gesture recognition, further emphasizing the value of attention mechanisms in improving model performance[14]. They also tested their model on the SHREC’17 dataset, yielding even higher results with 93.60% accuracy for 14 gestures and 90.70% for 28 gestures. The success of these models shows the effectiveness of leveraging attention mechanisms for sequential tasks such as gesture recognition. [15]

Recently, researchers have increasingly used Graph Convolutional Neural Networks (GCNN) for gesture recognition, taking advantage of the inherent structure in skeleton data to model the relationships between body joints as graphs. These methods have been applied with notable success, outperforming previous approaches that utilized traditional CNNs and RNNs for skeleton-based gesture recognition. [16]

Traditional gesture recognition systems relied heavily on feature extraction techniques, such as edge detection and motion tracking, to interpret hand gestures. For instance, methods such as Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT) were widely used for hand gesture recognition. These techniques achieved a certain level of success but faced challenges in handling occlusions, dynamic gestures, and variations in lighting conditions [17]. Moreover, traditional machine learning methods like Support Vector Machines (SVMs) and K-Nearest Neighbor (KNN) algorithms, though effective for static gestures, showed limitations in recognizing complex dynamic gestures that change over time [18].

The advent of deep learning techniques, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), has significantly improved the accuracy of hand gesture recognition. CNNs excel at spatial feature extraction, which is critical for image-based gesture recognition tasks [19]. For example, Ahmed et al. [20] developed a CNN-based model for recognizing Bangla hand gestures using a dataset of static images. Their system achieved a recognition accuracy of over 90%, demonstrating the effectiveness of deep learning in static gesture recognition.

However, for dynamic hand gestures, where the gesture involves a sequence of movements over time, Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) networks, have shown better performance. LSTMs can capture temporal dependencies, which are crucial for interpreting dynamic hand gestures. Some expert explored the use of a CNN-LSTM hybrid model for Bangla sign language recognition, incorporating both spatial and temporal information. Their approach achieved high accuracy, outperforming traditional machine learning methods.[21]

Another emerging trend is the use of attention mechanisms in hand gesture recognition models. Attention mechanisms allow the network to focus on the most relevant parts of the input, improving recognition performance in complex scenarios. Recent studies have shown that integrating attention mechanisms into CNN and RNN architectures can lead to better accuracy in both static and dynamic gesture recognition [22].

# **Methodology**

This research work aims to achieve the goal of classifying Bangla Sign Language hand gestures using various basic and advanced Neural Network models. For this research work we have apprehended a dataset from Kaggle as BDSL49, the following table describes the dataset.

**Table 1.** Description of Our Dataset Used

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Records** | **Labels** | **Resolution** | **Format** | **Participant** |
| 12,581 images | 0 to 37 for each alphabet | 224 \* 224 pixels | JPEG | 14 adults |

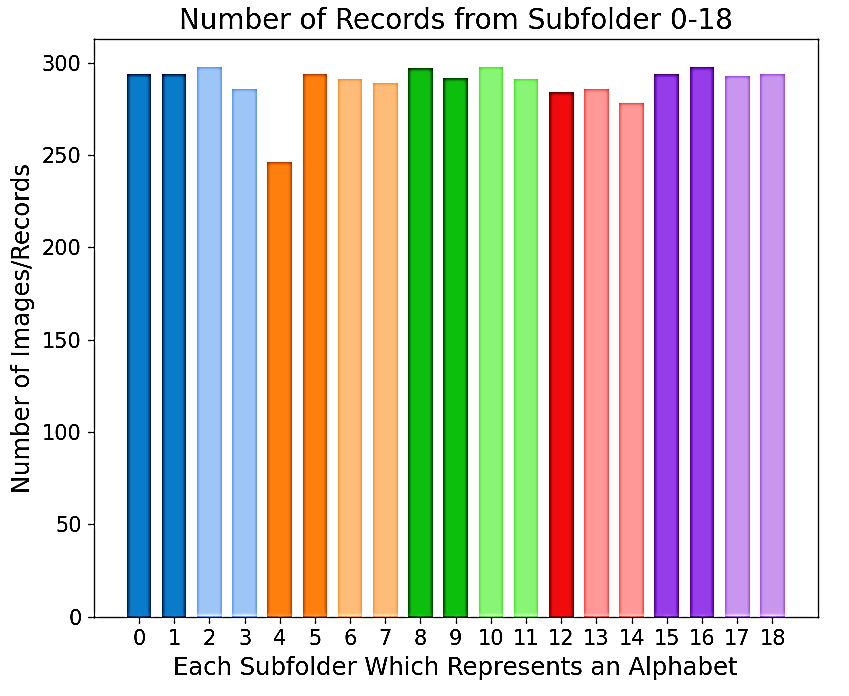
**Fig. 1.** Sample Data and Corresponding Bangla Alphabets

****

Our dataset occupies about 250 megabytes and is divided into a training set and a testing set split into 80:20 ratio, as such the training set has 11061 images and the testing set has 1520 images. The main subfolders have been divided each into 38 further subfolders named 0 to 37 each representing Bangla alphabets and containing the corresponding hand gesture images amalgamated from a participation of 14 volunteers.

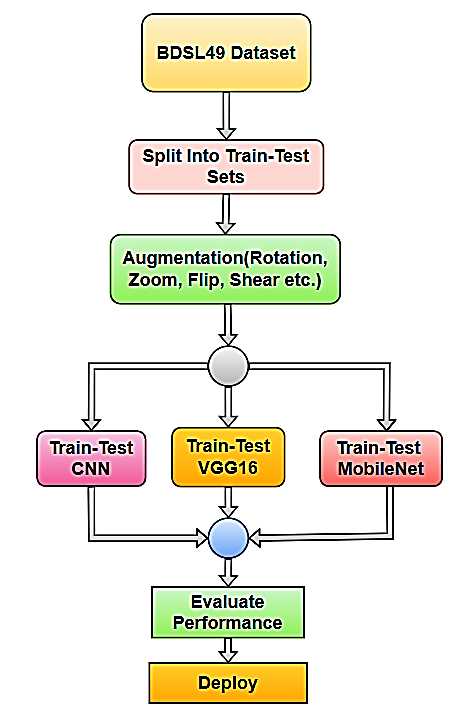
To get a proper understanding of our project setup we can have a look at the following table which outlines the experimental setup.

**Table 2.** Experimental Setup



|  |  |
| --- | --- |
| **Key Point** | **Description** |
| Aim | Classification of hand gesture images of Bangla Sign Language through various basic and advanced Neural Network models |
| Applications and Tools Used | Python 3.8, VScode, Numpy 1.24.4, TensorFlow 2.8, Scikit-Learn 1.3.2, Matplotlib 3.7.5 , Seaborn 0.13.2 etc. |
| Dataset Used | Bangla Sign Language Dataset (BDSL49) |
| Data Collection | Hand gesture images clicked using smartphone camera and compiled into a folder-structure |
| Participants | Hand images from 14 adult volunteer students from BUET |
| Procedure | Dataset Collection, Augmentation, Training various Neural Network models and Testing the models |
| Models Used | CNN, Pretrained VGG16 and Pretrained Mobilenet |
| Result Analysis | Analyzed through metrics such as accuracy histogram plot and confusion matrix |
| Ethical Concerns | The participants actively participated in the data collection process and all images were verified to make sure no regulations regarding privacy were violated |

To get a straightforward understanding of the workflow of our project we can take a look at the following flow diagram.

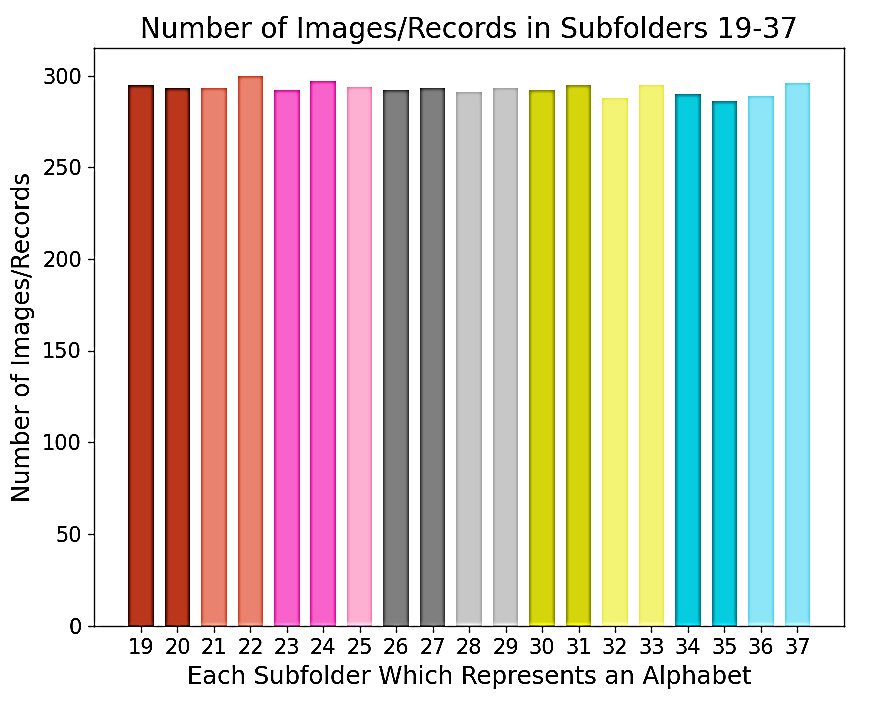


**Fig. 2**. Methodology of Our Work

# **IV. Results and Discussion**

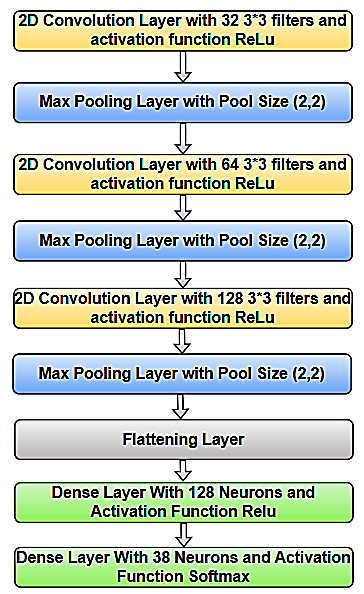
Initially we have plotted the class distribution to make sure the dataset is balanced which is a desirable quality when training any machine learning model.

**Fig. 3**. Class Distribution of BDSL49 Dataset(Class 0-18)



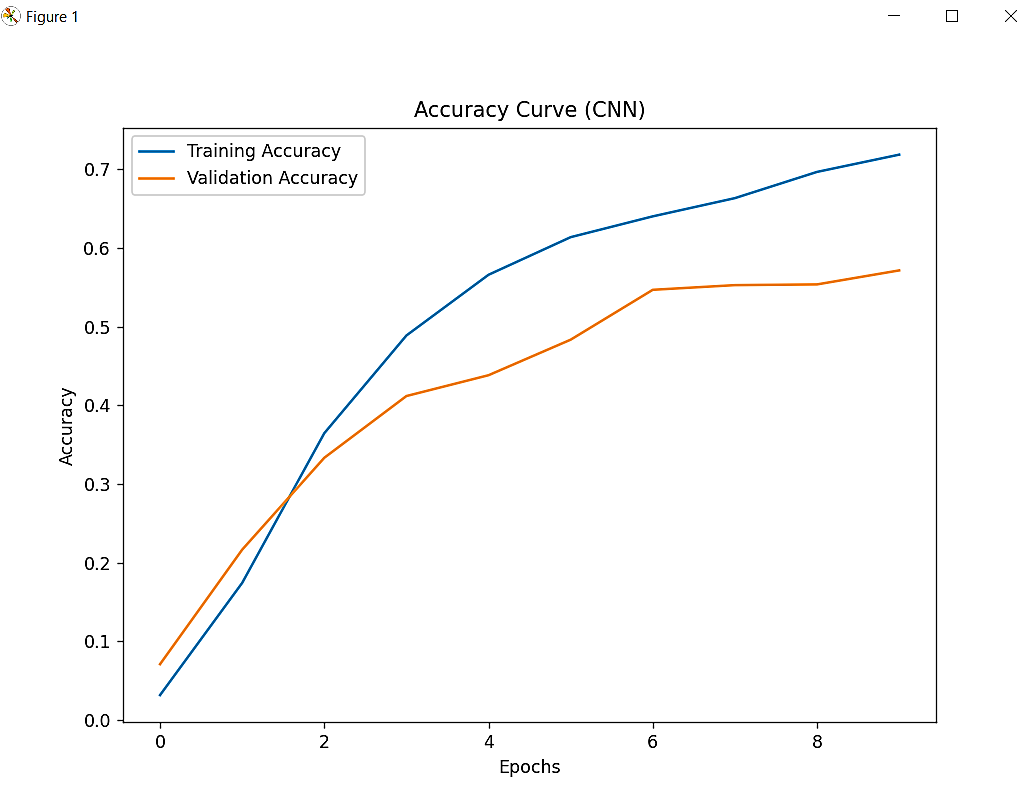
**Fig. 4**. Class Distribution of BDSL49 Dataset(Class 19-37)

From the plots it is apparent that the classes are fairly balanced so we haven’t implemented any oversampling or under-sampling techniques. For the evaluation purpose we have first stacked a CNN model.

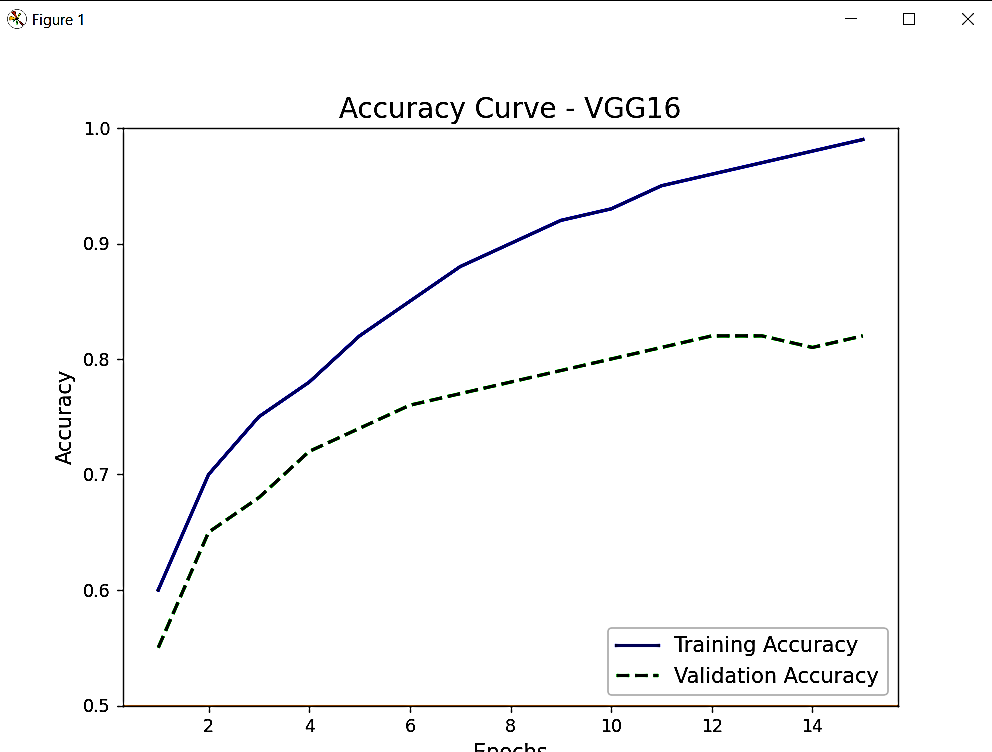


**Fig. 5**. Layer Architecture of Our CNN Model

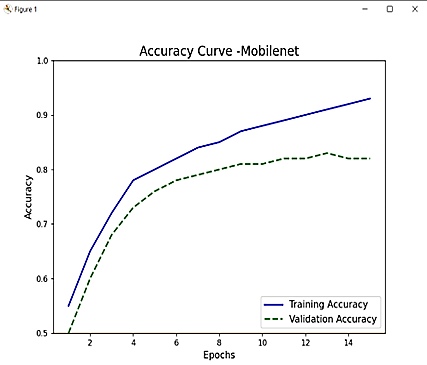
Using a CNN model as such along with Pretrained VGG16 and Mobilenet models we have done training and testing to get the following results.



**Fig. 5**. Accuracy Curve for Our CNN Model



**Fig. 6**. Accuracy Curve for Our VGG16 Model

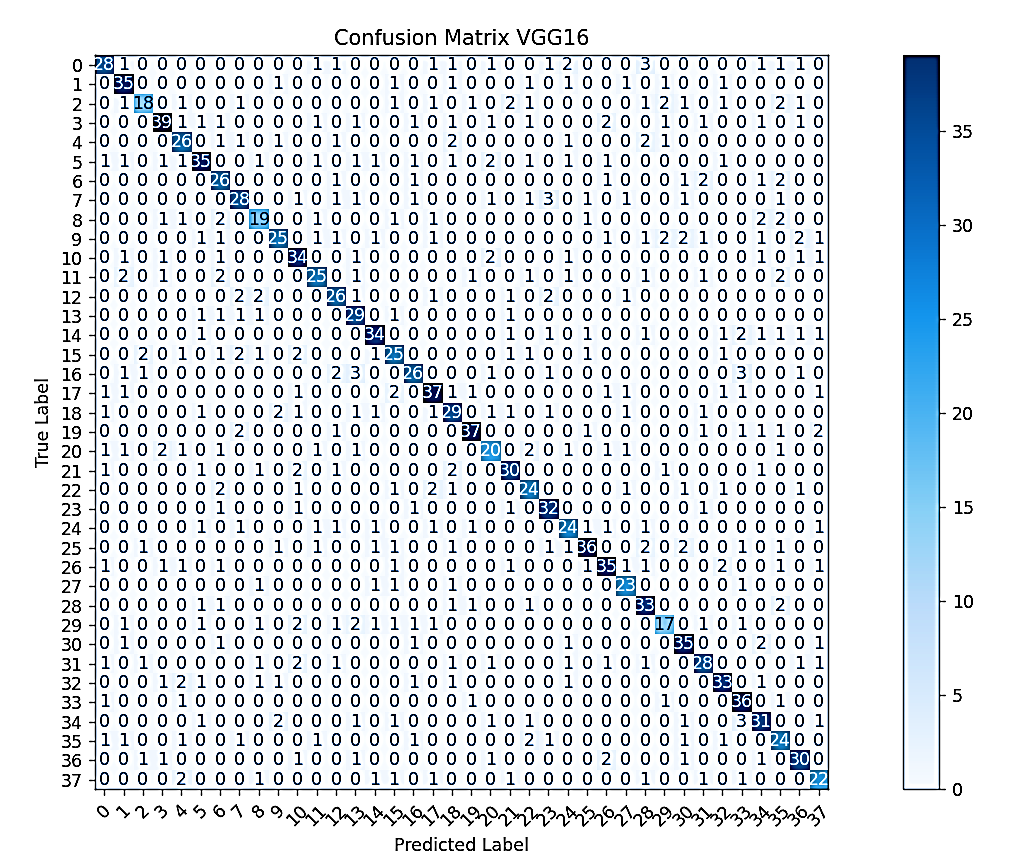


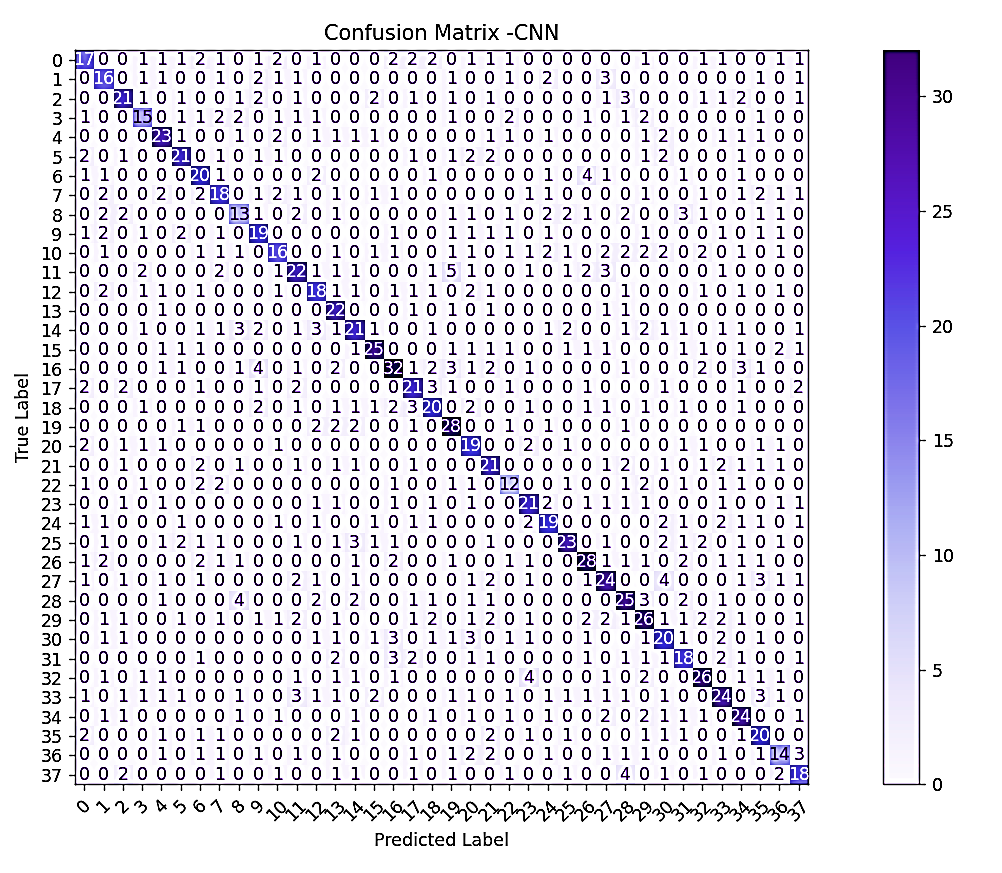
**Fig. 7.** Accuracy Curve for Our MobileNet Model

From evaluation we get 72% accuracy To make sure our models are not biased towards certain classes we have plotted

Confusion matrices for all 3 models as shown in the figures.

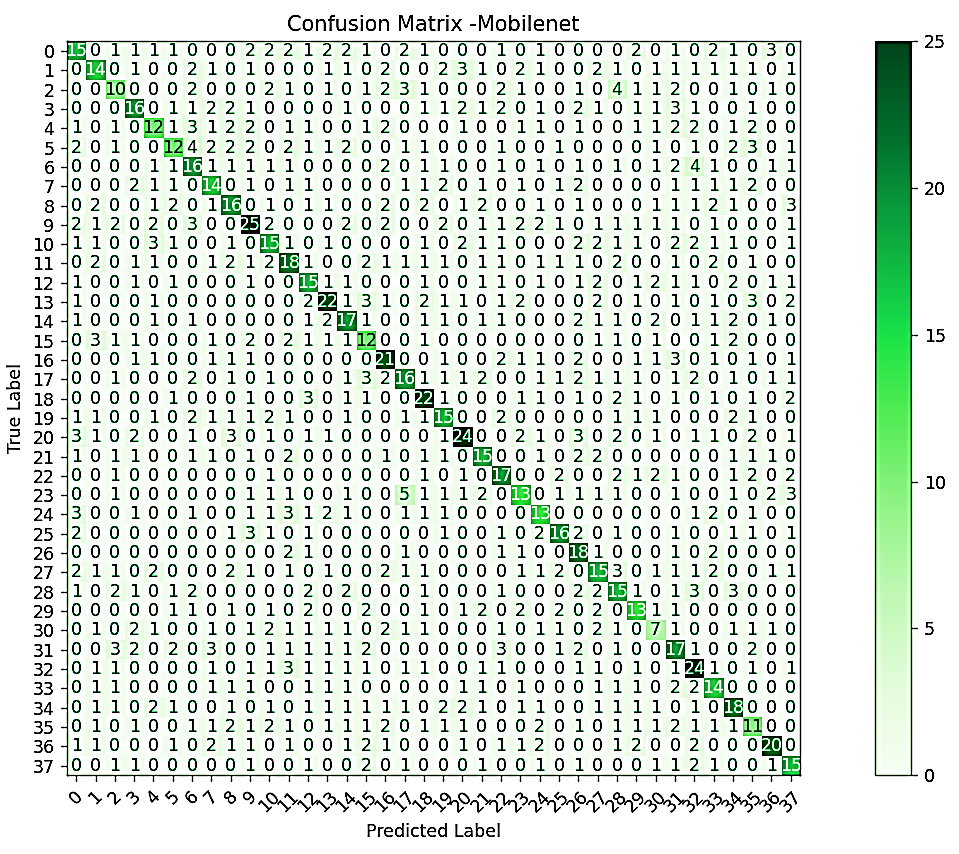
**Fig. 8.** Confusion Matrix for the CNN model





**Fig. 9.** Confusion Matrix for the VGG16 model

**Fig. 9.** Confusion Matrix for the MobileNet model



##### **V. Limitations**

From our works we have been able to achieve satisfactory results but better results can be achieved by using superior models and tuning the hyperparameters by trial and error.

##### **VI. Conclusion**

Ultimately, we can say that our work has been a success in terms of achieving satisfactory performance results for our models. We believe our work if improved can improve the social lives of Deaf-Mute individuals by aiding in their communication process.

##### **References**

[1] K. Neumann, S. Chadha, G. Tavartkiladze, X. Bu, and K. R. White, “Newborn and Infant Hearing Screening Facing Globally Growing Numbers of People Suffering from Disabling Hearing Loss,” International Journal of Neonatal Screening, vol. 5, no. 1, p. 7, Mar. 2019, doi: https://doi.org/10.3390/ijns5010007.

[2] J. M. Renauld and M. L. Basch, “Congenital Deafness and Recent Advances Towards Restoring Hearing Loss,” Current Protocols, vol. 1, no. 3, p. e76, Mar. 2021, doi: <https://doi.org/10.1002/cpz1.76>.

[3] Neha Singh, Neha Baranwal, and GC Nandi. Implementation and evaluation of dwt and mfcc based isl gesture recognition. In 2014 9th International Conference on Industrial and Information Systems (ICIIS), pages 1–7. IEEE, 2014.

[4] Kumud Tripathi, Neha Baranwal, and Gora Chand Nandi. Continuous dynamic Indian sign language gesture recognition with invariant backgrounds. In 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pages 2211–2216. IEEE, 2015.

[5] E. Ohn-Bar and M. M. Trivedi, ‘‘Joint angles similarities and HOG2 for action recognition,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, Jun. 2013, pp. 465–470.

[6]K. Thakkar and P. J Narayanan, ‘‘Part-based graph convolutional network for action recognition,’’ 2018, arXiv:1809.04983.

[7] S. Lu, D. Metaxas, D. Samaras, and J. Oliensis, ‘‘Using multiple cues for hand tracking and model refinement,’’ in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2003, p. 443.

[8] R. Vemulapalli, F. Arrate, and R. Chellappa, ‘‘Human action recognition by representing 3D skeletons as points in a lie group,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 588–595.

[9] J.Wang, Z. Liu, Y.Wu, and J. Yuan, ‘‘Mining actionlet ensemble for action recognition with depth cameras,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 1290–1297.

[10] C. Ma, A. Wang, G. Chen, and C. Xu, ‘‘Hand joints-based gesture recognition for noisy dataset using nested interval unscented Kalman filter with LSTM network,’’ Vis. Comput., vol. 34, nos. 6–8, pp. 1053–1063, Jun. 2018.

[11] X. Chen, G.Wang, H. Guo, C. Zhang, H.Wang, and L. Zhang, ‘‘MFA-Net: Motion feature augmented network for dynamic hand gesture recognition from skeletal data,’’ Sensors, vol. 19, no. 2, p. 239, Jan. 2019.

[12] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas, ‘‘CR-GAN: Learning complete representations for multi-view generation,’’ 2018, arXiv:1806.11191.

[13] J. Liu, A. Shahroudy, D. Xu, and G. Wang, ‘‘Spatio–temporal LSTM with trust gates for 3D human action recognition,’’ in Proc. Eur. Conf. Comput. Vis. Amsterdam, The Netherlands: Springer, 2016, pp. 816–833.

[14] S. Yan, Y. Xiong, and D. Lin, ‘‘Spatial temporal graph convolutional networks for skeleton-based action recognition,’’ in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 7444–7452.

[15] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, ‘‘Skeleton-based action recognition with spatial reasoning and temporal stack learning,’’ in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 103–118.

[16] Y. Chen, L. Zhao, X. Peng, J. Yuan, and D. N. Metaxas, ‘‘Construct dynamic graphs for hand gesture recognition via spatial–temporal attention,’’ 2019, arXiv:1907.08871.

[17] Pradhan, A., & Sethi, S. (2014). A Survey on Gesture Recognition Systems. *International Journal of Computer Applications*, 104(15), 23-26.

[18] Raheja, J. L., & Chaudhary, A. (2015). Real-time hand gesture recognition using SIFT. *International Journal of Computer Applications*, 69(1), 37-43.

[19] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[20] Ahmed, S., Hasan, M. M., & Khan, M. R. (2018). Bangla Sign Language Recognition Using Convolutional Neural Network. *Proceedings of the International Conference on Informatics, Electronics, and Vision (ICIEV)*, 1-6.

[21] Chowdhury, M. S., & Islam, M. N. (2019). A Hybrid CNN-LSTM Model for Dynamic Bangla Sign Language Recognition. *International Journal of Machine Learning and Computing*, 9(2), 190-196.

[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 5998-6008.