[1]Rajesh Yakkundimath

[2]Ramesh Badiger

[3]Naveen Malvade

"Deep Learning Based Classification of Single-Hand South Indian Sign Language Gestures"

Abstract: *-* Gesture recognition is a branch of computer science and language technology dedicated to utilizing mathematical algorithms for the analysis of human gestures. Within the realm of non-verbal communication, the pivotal role of human arm movements and gestures remains a focal point. This research introduces advanced multi-stream deep transfer learning models tailored for identifying signs from South Indian languages, specifically Kannada, Tamil, and Telugu. The primary aim is to offer support to individuals encountering speech disorders or disabilities. The key deep transfer learning models utilized include Inception-V3, VGG-16, and ResNet-50, which have been modified and improved to attain heightened classification efficacy. The dataset comprises 35,000 images capturing single-hand gestures. In the realm of models, Inception-V3 exhibits the utmost test accuracy, standing at 91.45%, alongside a validation accuracy of 93.45% when tasked with the classification of single-hand gesture images across thirty-five distinct categories. The importance of this study lies in its prospective utility for creating an automated system that can support and improve the functional capabilities of individuals facing speech disorders or disabilities.

Keywords: Inception-V3, Deep transfer learning, South Indian languages, Speech disorders, Classification efficacy, Automated assistance system

I.INTRODUCTION

The emergence of Artificial Intelligence (AI) technologies has introduced substantial possibilities to develop solutions that bring about positive transforma tions in the World Health Organization reports that around 15% of the global population experiences the challenges associated with disabilities. This demographic faces unique circumstances, and understanding their lives is crucial for fostering inclusivity and providing appropriate support totaling 1.2 billion individuals, who face various forms of disability. In the pursuit of inclusivity, a myriad of entities, including corporations, institutions, and individuals, have harnessed the power of AI tools to aid in daily tasks the research landscape in machine learning and artificial intelligence has gradually pivoted towards the seamless integration of accessibility and disability-related advancements.

This evolution manifests in a dual focus: one on aiding medical treatments for disabilities, and the other on augmenting accommodations for individuals facing diverse disabilities.

A noteworthy area where AI has made substantial strides is in facilitating Communication among individuals with disabilities often relies on non-verbal cues, such as the interpretation of facial expressions, gestures, sign languages, and various forms of body language. Sign language, a method primarily employed by deaf or hearing-impaired individuals, relies on hand gestures and movements instead of verbal communication. Gesture and sign language recognition have emerged as dynamic topics within the realms of computer vision and pattern recognition. Despite recent advancements, achieving swift and robust hand gesture recognition remains an ongoing challenge.

[1] Department of Computer Science and Engineering, K. L. E. Institute of Technology, Visvesvaraya Technological University, Belagavi 590018, Karnataka, India

[2]Department of Computer Science and Engineering, Tontadarya College of Engineering, Visvesvaraya Technological University, Belagavi 590018, Karnataka, India

Gadag 582 101, Karnataka, India

[3]Department of Information Science and Engineering, S.K.S.V.M. A. College of Engineering and Technology, Lakshmeshwar 582116, Karnataka, India

*Corresponding Author: Rajesh Yakkundimath

*Department of Computer Science and Engineering, K. L. E. Institute of Technology, Visvesvaraya Technological University, Belagavi 590018, Karnataka, India

Figure 1 illustrates examples of hand gesture images, and Table 1 furnishes label encoding specifics for distinct single-hand gestures in addition to their corresponding characters, the abstract encompasses the rich variety of South Indian languages. In order to acquire a comprehensive understanding of the contemporary advancements in computer vision and pattern recognition applications, an exhaustive literature survey was undertaken. The abridged summary of this survey is provided below to encapsulate the prevailing state-of-the-art developments.

Figure 1: Various single-hand gesture images

Table 1. Label encoding/numerical code assignments for the considered single hand gesture

Class Label (Numerical Code)	Letter in Kannada/Telugu/Tamil/ English Language	Class Label (Numerical Code)	Letter in Kannada/Telugu/Tamil/ English Language
1	ಕ/ క/ க /Ka	19	ಧ / ధ /-/Dha
2	ಖ/ఖ/ கா/ Kha	20	ನ/న/ண/Na
3	ಗ/ గ / ங/ga	21	ಪ/ ప/ ப/Pa
4	ಘ /ఘ / ஙா/Gha	22	ಫ/ఫ/ப/Pha
5	ಙ/ ఙ / ஞ/Na	23	ಬ / బ/- / Ba
6	ಚ /చ / ச/Ca	24	ಭ / భ /-/Bha
7	ಛ / ఛ / சா/Cha	25	ಮ/మ /ம/Ma
8	ಜ / జ / ஜ/Ja	26	ಯ /య/ய/Ya
9	ಝ /ఝ/ ஜா/Jha	27	ರ / ర / ர/ Ra
10	ಞ/ ఞ/ ண/Na	28	ಲ / ల / ல/ La
11	ಟ / ట / ட/Ta	29	ವ / వ / வ/Va
12	ಠ /ఠ/ ட/Tha	30	ಶ/ శ / ஶ/ Sha
13	ಡ / డ /-/ Da	31	ಷ/ ష/-/Sha
14	ಢ/ఢ /-/ Dha	32	ಸ/ స/ ஷ/Sa
15	ಣ/ణ/ ண/Na	33	ಹ/ హ/ ஹ/ha
16	ತ/త / த/Ta	34	ಳ/ ళ/-/La
17	ಥ/థ/தா/Tha	35	ಕ್ಷ/ క్ష / க்ஷ/ksha
18	ದ /ద /-/Da		

Abualkishik et al. (2023) presented a groundbreaking approach to gesture recognition, employing a fusion of convolutional neural networks (CNN) and Internet of Things (IoT) technology. The dataset consists of 62,000 images (64x64 pixels) portraying the 30 letters of the alphabet and a single word. Notably, the devised system attains a remarkable accuracy of 99.8%. Alam et al. (2022) conducted an extensive review of related works on hand gesture recognition and fingertip detection, covering both traditional computer vision methods and deep learning-based approaches. Al Farid et al. (2022) presented a systematic review of vision-based hand gesture recognition systems, emphasizing their significance in applications like human-computer interaction and sign language recognition. Their review included discussions on the challenges associated with recognizing hand gestures, reporting an average identification accuracy of 86.6%.

Alyami et al. (2023) presented a successful application of a transformer-based model with landmark key points for the recognition of isolated sign language. In their evaluation on the KArSL-100 dataset, covering Arabic and Argentinian sign languages, the proposed pose-based transformer exhibited the highest accuracy rates. Specifically, it achieved 99.74% and 68.2% accuracy in signer-dependent and independent modes, respectively. In a separate investigation, Anami and Bhandage (2019) developed a conventional image processing approach for classifying Bharatanatyam mudra images. Their methodology included the utilization of feature extraction techniques such as Hu-moments, Eigenvalues, and vertical-horizontal intersection. Depending on the specific method used, the overall classification accuracy ranged from 94.71% to 98%.

Bora et al. (2023) introduced an innovative real-time recognition strategy for Assamese sign language, employing MediaPipe and deep learning techniques. The model utilized a convolutional neural network (CNN) architecture for gesture classification, achieving an impressive accuracy rate of 99%. Chakraborty et al. (2023) presented a distinct approach to sign language gesture recognition, utilizing deep learning techniques, particularly a CNN-based architecture for landmark detection and tracking key points on the hand. Damaneh et al. (2023) proposed a novel structure that combines a CNN with classical non-intelligent feature extraction methods, such as the ORB descriptor and Gabor filter, to identify static hand gestures in sign language. This suggested structure demonstrated the highest accuracy of 99.92% for American Sign Language.

In the study by Das et al. (2023), they proposed an integrated framework for the identification of Bangla Sign Language, which combines a deep transfer learning-based convolutional neural network with a random forest classifier. The system demonstrated impressive performance in character recognition, achieving metrics of 91.67%

accuracy, 93.64% precision, 91.67% recall, and a 91.47% f1-score. Additionally, the model excelled in digit recognition, with accuracy, precision, recall, and f1-score values of 97.33%, 97.89%, 97.33%, and 97.37%, respectively.

In a distinct inquiry, De Castro et al. (2023) undertook an examination delving into a multi-stream deep learning model crafted for discerning signs in Brazilian, Indian, and Korean Sign Languages. Their approach encompassed the employment of single-stream and multi-stream 3D Convolutional Neural Networks, yielding an accuracy of 0.91 ± 0.07 and an f1-score of 0.90 ± 0.08. Mia et al. (2023) introduced an inventive approach to sign language recognition, underscoring the importance of hidden transfer learning in enhancing communication for individuals with hearing and speech disabilities. Simultaneously, Pandey et al. (2023) presented a groundbreaking prototype featuring a Feed Forward Neural Network (FFNN) for automatic sign language recognition. Their goal was to facilitate communication for both deaf individuals and the general public. Additionally, they integrated gesture recognition with a voice processing system employing the Hidden Markov Model (HMM).

Rajalakshmi and collaborators (2023) introduce a narrative-focused hybrid deep neural network crafted for recognizing sign gestures in both Indian and Russian settings. Their aim is to establish a comprehensive framework adept at tracking and extracting multi-semantic properties, covering non-manual elements and manual co-articulations. Concurrently, Sudhakar and colleagues (2023) explore the identification and acknowledgment of sign language gestures, emphasizing the importance of sign language recognition in enhancing communication accessibility for individuals who are deaf and hard-of-hearing.

Tran et al. (2020) introduce a framework specifically developed for the prompt spotting and identification of hand gestures. This system integrates an RGB-D camera with a 3D convolutional neural network to enable instantaneous recognition. The authors emphasize the significance of hand gesture recognition, particularly in domains like human-robot interaction. They detail the obstacles linked to achieving real-time detection and identification of hand gestures.

In their 2020 study, Wang and colleagues devised a supervised convolutional neural network tailored for the recognition of human hand gestures in K–12 classrooms employing a double-teacher instruction mode. The architecture of their model, encompassing three convolutional layers designed to extract features from infrared hand images, attains an impressive overall accuracy in classification, reaching 92%.

Xu and colleagues (2023) present an innovative online lightweight framework designed for dynamic gesture recognition in untrimmed videos. The framework is complemented by a unique gesture dataset named ZJU (Zhejiang University) gesture. The significance of dynamic gesture recognition is underscored by the authors in diverse applications, including human-computer interaction and surveillance. Their focus centers on addressing the difficulties associated with identifying gestures within untrimmed videos.

Zengeler et al. (2018) introduce a framework designed for hand gesture recognition within the context of human-machine interaction (HMI) in automotive settings. The authors emphasize the significance of this technology in enhancing safety and minimizing driver distraction. They highlight the difficulties related to identifying and interpreting hand gestures in realistic environments.

The review of existing literature reveals substantial research in the domain of sign language and hand gesture recognition, often employing deep learning techniques for human pose and scene identification from images. Notably, when focusing on single-hand gestures, specifically within the context of South Indian Sign Language (SISL), a noticeable gap exists in the literature. This void serves as the primary motivation for the current study. The paper is structured into four sections, where Section 2 outlines the proposed methodology, Section 3 presents the experimental results, and Section 4 offers the concluding remarks.

II.PROPOSED METHODOLOGY

The suggested approach comprises two primary phases: the initial one involves the preparation of an image dataset, followed by the second stage which encompasses deep learning-based classification. Figure 2 depicts a block diagram that elucidates the successive stages integral to the proposed methodology.
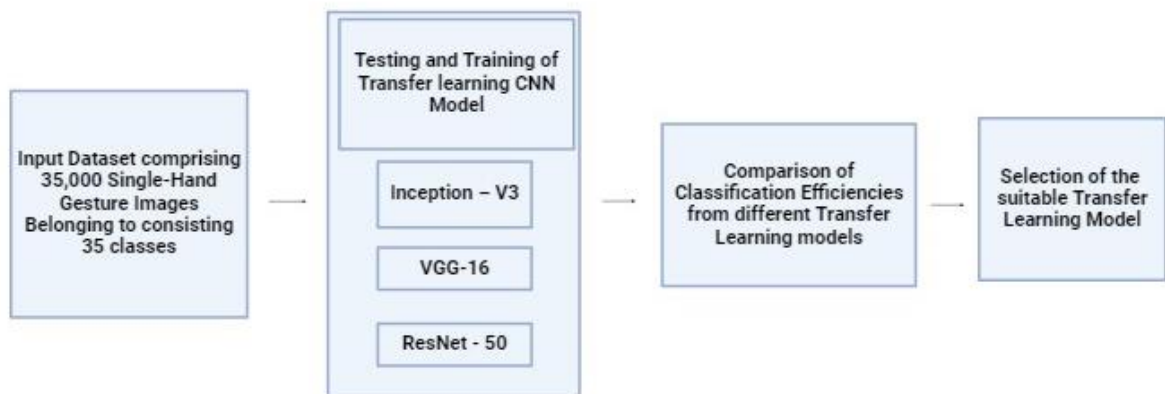
Figure 2: Block diagram of the proposed methodology

Dataset Preparation

The dataset preparation process meticulously considered 35 distinct categories of single-hand gestures. Images were captured under natural light conditions using a Nikon D3300 digital SLR camera, renowned for its impressive 24.2-megapixel resolution. Each image, measuring 1080 x 2400 pixels, played a pivotal role in contributing to an extensive original dataset that consisted of 17,500 images. Notably, these images were evenly distributed, with each single-hand gesture category comprising 500 images.

To enrich the dataset and enhance its diversity, various augmentation techniques were employed. These techniques included translation, arbitrary rotations, shearing, scaling, and flipping, resulting in a substantial augmentation that doubled the dataset's size to an impressive 35,000 images. This augmentation process aimed to provide the model with a robust and versatile training set, capable of handling a wide range of scenarios and variations in hand gestures.

Recognizing the importance of computational efficiency and storage considerations, all images, initially of varying sizes, underwent a resizing procedure. The resizing was carried out uniformly, with the images being adjusted to a standardized 300 x 300 pixels. This standardization not only reduced computational processing time but also helped streamline storage requirements, ensuring a more efficient and manageable dataset for subsequent analyses.

The final dataset underwent a systematic partitioning, a crucial step in preparing it for training, validation, and testing phases. Specifically, 70% of the images (24,500) were allocated for training, allowing the model to learn from the majority of the data. Another 15% (5,250 images) were set aside for validation, serving as an independent dataset to fine-tune hyperparameters and prevent overfitting. The remaining 15% (5,250 images) constituted the testing set, enabling the assessment of the model's generalization performance on unseen data.

Table 1, presented in the study, offers a comprehensive overview of the refined image dataset. This table serves as a valuable reference, providing detailed insights into the distribution of images across categories and their respective allocation for training, validation, and testing. Such transparency in dataset details is vital for the reproducibility and understanding of the study's results, allowing researchers to assess and build upon the work presented.

In conclusion, the dataset preparation process outlined above demonstrates a meticulous and thoughtful approach to constructing a robust dataset for training a model on single-hand gestures. The incorporation of diverse augmentation techniques, standardization of image sizes, and systematic partitioning for training and evaluation collectively contribute to the quality and reliability of the dataset, setting the stage for meaningful and impactful research in the domain of gesture recognition.

II.1   CNN Classifiers

In the present investigation, the recognition and categorization of images depicting single-hand gestures are executed utilizing three well-established transfer learning Convolutional Neural Network (CNN) models: The Inception-V3 model, as proposed by Szegedy et al. in 2016, VGG-16, and ResNet-50 by He et al. in the same year,

are extensively detailed in the literature by Too et al. (2019). Algorithm 1 precisely outlines the procedural steps for training these specified CNN models on a carefully curated dataset.

Algorithm 1. Training and assessing pre-trained CNN models on the dataset of single-hand gesture images.

Input: Data-set of single hand gesture images.

Output: Recognized pictures..

Commence

Step 1. Incorporate the dataset of images encompassing single hand gestures.

Step 2. Formulate batches of datasets for training and testing for images.

Step 3. Load the foundational CNN model and tailor solely the concluding layer.

Step 4. Assemble and train the foundational CNN model with diverse dropout rates, as well as various optimizers and activation functions.

Cease."

<div align="center">III.EXPERIMENTAL RESULT AND DISCUSSIONS</div>

The experimentation on single-hand gesture classification is executed through the utilization of the Deep Learning Toolbox provided by the MATLAB R2022b programming platform. Transfer learning is facilitated by importing and preparing pre-trained CNN models, wherein suitable layer properties are edited using the Deep Network Designer application. In the process, the last learnable layer and the output or classification layer in all models are replaced to align with the classes in the newly constructed single-hand gesture image dataset. The entire CNN model training and testing operations are carried out on a singular workstation running the Windows 11 operating system. This workstation is configured with an Intel Core i9–13,900 processor, 16 GB of RAM, and an NVIDIA GPU boasting 12 GB of memory.

To exercise control over model training, pivotal training options such as initial learning rate, validation frequency, number of epochs, and mini-batch size are initialized to 0.0001, 10, 30, and 35, respectively. The ReLu function is employed to activate all hidden layers, while the softmax function is applied for the activation of the output layer. The network undergoes fine-tuning through the utilization of a stochastic gradient descent (SGD) algorithm, complemented by a categorical cross-entropy logarithmic loss function. In an effort to enhance overall accuracy and expedite the learning process, batch normalization is implemented.

The augmented image dataset was utilized to train and validate customized CNN models. The training progress, validation accuracy, loss, and other training details for each CNN model are visually represented in Figs. 3–5. Figs. 6–8 illustrate the confusion matrices of the CNN models when assessed with the test image dataset. To compare the performance of the CNN models, evaluation metrics such as precision, recall, and F1 scores, derived from the confusion matrices, were employed. Table 1 presents the average evaluation metric scores for all 35 classes of single-hand gesture images across the considered CNN models. According to the results in Table 1, the Inception-V3 model exhibits the most superior performance, followed by the ResNet-50 and VGG-16 models, in that sequence. Notably, the Inception-V3 model achieved the highest evaluation metrics, along with maximum validation and testing accuracies of 93.45% and 94.10%, respectively. Graphical representation of the performance comparison results for all considered pre-trained CNN models is provided in Fig. 9.
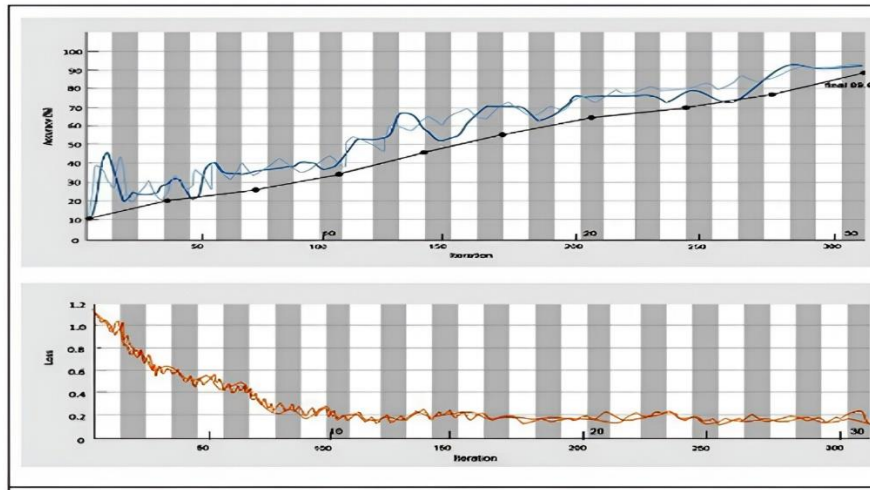
Figure 4: Plot of validation accuracy and loss graph while training Inception CNN model
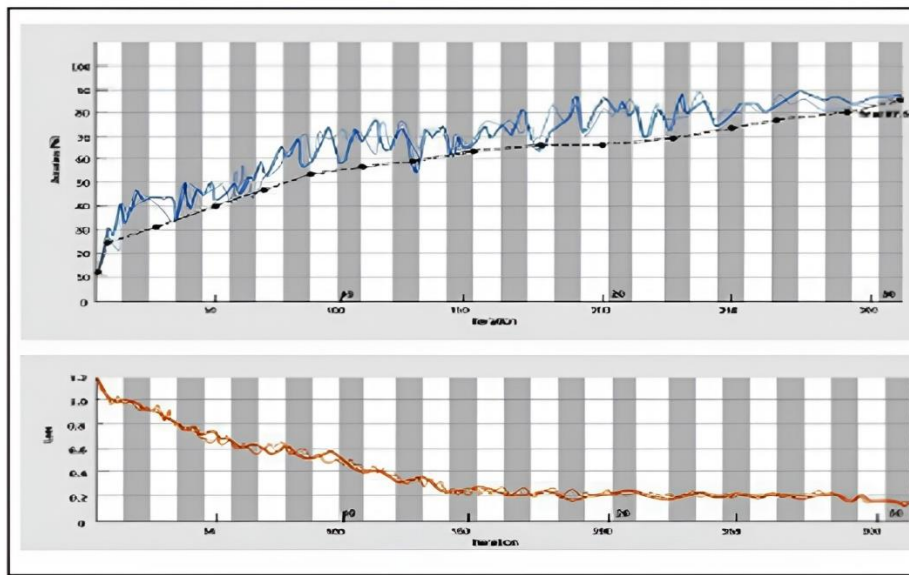

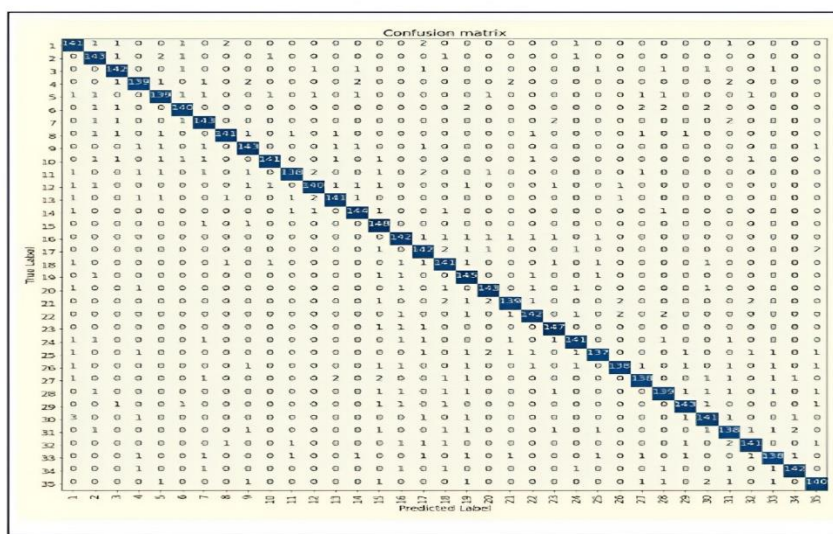**Figure 5: Plot of validation accuracy and loss graph while training Inception CNN model**

Figure 6: Confusion matrix plotted on the test dataset for the trained inception- V3 CNN model.
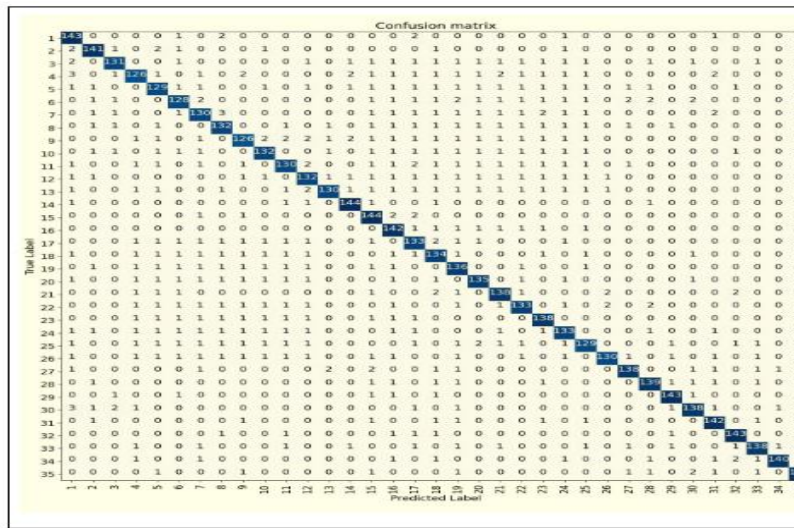
Figure 7: Confusion matrix plotted on the test dataset for the trained ResNet-50 CNN model.
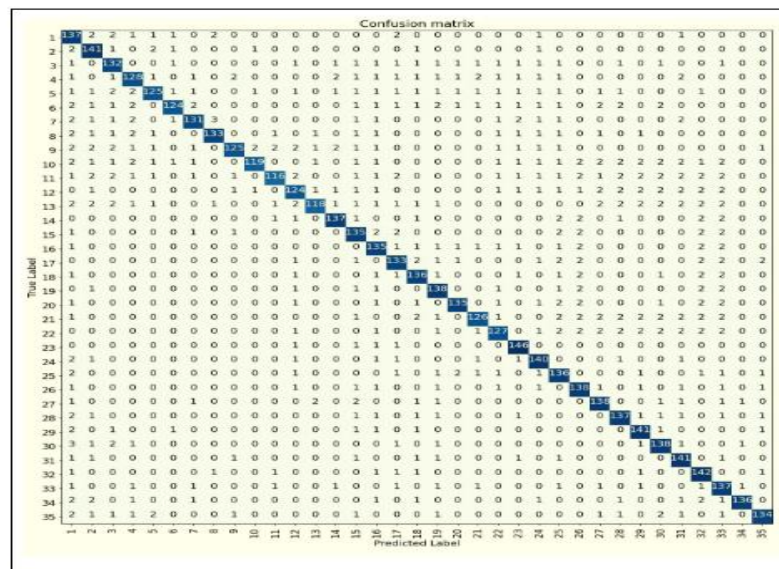

Figure 8: Confusion matrix plotted on the test dataset for the trained Inception-V3 CNN model.

Table 1. Evaluation metrics derived from the confusion matrices plotted for the considered CNN models.

Sl. No	CNN model	Average Performance metrics across all the single-hand gesture classes			Average Validation Accuracy (%)	Average Test Accuracy (%)
		Precision	Recall	Fl Score		
1	Inception-V3	0.9411	0.9426	0.9394	93.45%	94.1%
2	ResNet-50	0.9031	0.9000	0.9009	89.68%	90.3%
3	VGG-16	0.8883	0.8929	0.8877	87.59%	88.7%
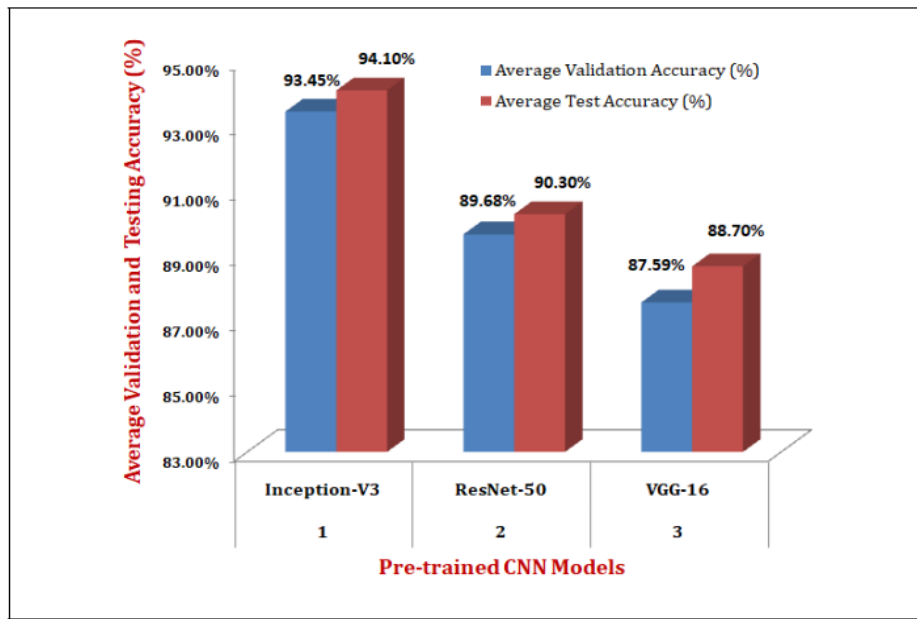
Figure 9: Performance evaluation results of all the pre trained CNN models

IV.CONCLUSION

In this current study, we utilized leading commercially available pre-existing convolutional neural network (CNN) architectures, namely Inception-V3, ResNet-50, and VGG-16, which are utilized for the categorization of single-hand gesture images encompassing 35 distinct classes. The models underwent slight customization to accommodate the number of image classes and were fine-tuned to enhance classification performance. While all CNN models demonstrated impressive results, Inception-V3 notably surpassed ResNet-50 and VGG-16, achieving an average classification accuracy of 94.45% in the task. The applied significance of this study resides in creating an automated system capable of recognizing gestures in the South Indian language, employing both static images and live streaming videos. Future directions include deploying the latest CNN models with an augmented image dataset, incorporating publicly available data, to further enhance the scope of this work.

Acknowledgement

Declaration of Competing Interest

The corresponding author certifies that there is no conflict of interest.

Funding

REFERENCES

[1]   Abualkishik, A., Alzyadat, W., Al Share, M., Al-Khaifi, S. and Nazari, M., 2023. Intelligent Gesture Recognition System for Deaf People by using CNN and IoT. Int. J. Advance Soft Compu. Appl, 15(1), pp.144-158.

[2]   Alam, M.M., Islam, M.T. and Rahman, S.M., 2022. Unified learning approach for egocentric hand gesture recognition and fingertip detection. Pattern Recognition, 121, pp.108-200.

[3]   Al Farid, F., Hashim, N., Abdullah, J., Bhuiyan, M.R., Shahida Mohd Isa, W.N., Uddin, J., Haque, M.A. and Husen, M.N., 2022. A structured and methodological review on vision-based hand gesture recognition system. Journal of Imaging, 8(6), pp.1-19.

[4]   Alyami, S., Luqman, H. and Hammoudeh, M., 2023. Isolated Arabic Sign Language Recognition Using A Transformer-based Model and Landmark Keypoints. ACM Transactions on Asian and Low-Resource Language Information Processing. https://doi.org/10.1145/3584984.pp.1-19.

[5] Anami, B.S. and Bhandage, V.A., 2019. A comparative study of suitability of certain features in classification of bharatanatyam mudra images using artificial neural network. Neural Processing Letters, 50(1), pp.741-769.

[6] Bora, J., Dehingia, S., Boruah, A., Chetia, A.A. and Gogoi, D., 2023. Real-time Assamese Sign Language Recognition using MediaPipe and Deep Learning. Procedia Computer Science, 218, pp.1384-1393.

[7] Chakraborty, S., Sarkar, S., Paul, P., Bhattacharjee, S. and Chakraborty, A., 2023. Sign language recognition using landmark detection, GRU and LSTM. American Journal of Electronics & Communication, 3(3), pp.20-26.

[8] Damaneh, M.M., Mohanna, F. and Jafari, P., 2023. Static hand gesture recognition in sign language based on convolutional neural network with feature extraction method using ORB descriptor and Gabor filter. Expert Systems with Applications, 211, pp.118559.

[9] Das, S., Imtiaz, M.S., Neom, N.H., Siddique, N. and Wang, H., 2023. A hybrid approach for Bangla sign language recognition using deep transfer learning model with random forest classifier. Expert Systems with Applications, 213, pp.118914.

[10] De Castro, G.Z., Guerra, R.R. and Guimarães, F.G., 2023. Automatic translation of sign language with multi-stream 3D CNN and generation of artificial depth maps. Expert Systems with Applications, 215, pp.119394.

[11] He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.48550/arXiv.1512.03385, 7(3), pp. 171–180.

[12] Mia, K., Islam, T., Assaduzzaman, M., Shaha, S.P., Saha, A., Razzak, M.A., Dhar, A. and Sarker, T., 2023. Isolated sign language recognition using hidden transfer learning. International Journal of Computer Science and Information Technology Research, 11(1), pp.28-38.

[13] Pandey, A., Chauhan, A. and Gupta, A., 2023. Voice Based Sign Language Detection For Dumb People Communication Using Machine Learning. Journal of Pharmaceutical Negative Results, 14(2), pp.22-30.

[14] Rajalakshmi, E., Elakkiya, R., Subramaniyaswamy, V., Alexey, L.P., Mikhail, G., Bakaev, M., Kotecha, K., Gabralla, L.A. and Abraham, A., 2023. Multi-Semantic Discriminative Feature Learning for Sign Gesture Recognition Using Hybrid Deep Neural Architecture. IEEE Access, 11, pp.2226-2238.

[15] Sudhakar, R., Gayathri, V., Gomathi, P., Renuka, S. and Hemalatha, N., 2023. Sign Language Detection. South Asian Journal of Engineering and Technology, 13(1), pp.49-56.

[16] Szegedy, C., Liu, W., Jia, Y., Scott Reed, P,S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. 2014. Going Deeper with Convolutions.2015 IEEE Conference on Computer Vision and Pattern Recognition, https://doi.org/10.48550/arXiv.1409.4842, pp.1-12.

[17] Tran, D.S., Ho, N.H., Yang, H.J., Baek, E.T., Kim, S.H. and Lee, G., 2020. Real-time hand gesture spotting and recognition using RGB-D camera and 3D convolutional neural network, Applied Sciences, 10(2), pp.722.

[18] Too E C, Yujiana Li, Njukia S, Yingchun L.2019. A comparative study of fine-tuning deep learning models for plant disease identification. Computers and Electronics in Agriculture, 161, pp.272-279.

[19] Wang, J., Liu, T. and Wang, X., 2020. Human hand gesture recognition with convolutional neural networks for K-12 double-teachers instruction mode classroom. Infrared Physics & Technology, 111, pp.103464.

[20] Xu, C., Wu, X., Wang, M., Qiu, F., Liu, Y. and Ren, J., 2023. Improving dynamic gesture recognition in untrimmed videos by an online lightweight framework and a new gesture dataset ZJUGesture. Neurocomputing, 523, pp.58-68.

[21] Zengeler, N., Kopinski, T. and Handmann, U., 2018. Hand gesture recognition in automotive human–machine interaction using depth cameras. Sensors, 19(1), pp.59.