Identification of Hand Gestures for Sign Language Interpretation Through LSTM and GRU

S.Pooja *

ABSTRACT

In the past, people who have hearing disability have been neglected and they unable to access the resources that would enable them to communicate effectively. The developments in contemporary technology contain wide range of instruments and applications, which have been created with the intention of enhancing the lives of those who are hard of hearing. This work uses four machine learning algorithms designed by thinking to identify hand movements for the American Sign Language (ASL) alphabet. It is a thorough investigation. The main goal of this research is to use modern methods to eradicate the communication gap between those who have hearing disability and others who do not. The models used in this study, which included two-layer LSTM and GRU, GRU then LSTM, and LSTM then GRU, were trained and evaluated using a large dataset that included more than 87,000 photos of hand motions associated with the ASL alphabet. Extensive studies were carried out whereby the models' architectural design characteristics were altered in order to get the highest possible identification accuracy. Our study's experimental findings showed that, out of all the models, GRU and LSTM combined to reach an amazing accuracy rate of 99.04%. Two-layer GRU produced an accuracy rate of 98.56%, two-layer LSTM produced the lowest accuracy of 98.56%, and LSTM followed GRU in achieving an accuracy rate of 98.78%.

Keywords: Sign Language, Sign Language Recognition, Accessibility Aids, Deep Learning, Computer Vision.

1. INTRODUCTION

Humans have used a wide range of communication methods from past century, such as speaking, writing, gesturing, and making noises. However, sign language serves as the main form of communication and social contact for those who are deaf or have hearing difficulties. It overcomes the limitations posed by hearing loss, which severely restricts spoken communication. People with these disorders have limited developmental possibilities due to communication barriers. Body language, facial emotions, and manual gestures are used to communicate meaning and messages in sign language, an organic non-verbal language. While people use sign language differently in different countries, some signals may be comparable in certain ways. Regretfully, there isn't a universal way for all individuals with hearing loss to communicate worldwide [1].

According to the World Health Organization's (WHO) latest research, more than one billion individuals are susceptible to hearing impairment from extended exposure to loud sounds, and 432 million grown people and 34 million children worldwide suffer from serious auditory impairment [2]. Indeed, people who lack in ability to hear in any setting find no way to talk. Due to the prevalence of hearing impairment and auditory problems, an abundance of investigators and programmers in speech identification as well as other interdisciplinary industries have found themselves attracted to the field to carry out their studies with the goal to assist those who suffer from auditory impairments. Their objective is to improve the quality of life of individuals experiencing auditory impairments by facilitating conversation as well as social interactions. Therefore, developing a deep learning technology-based sign language recognition system is essential for deaf people to understand ordinary people's sign language and vice versa, given the continuously expanding deaf population. This method would facilitate communication between the deaf community and the general public. Consequently, individuals with hearing impairments will have the chance to connect socially and build relationships by being more involved in society [3,4]. These days,

Department of Artificial Intelligence and Data Science
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, Inida
* Corresponding Author

translators are the only way that persons with hearing loss may communicate with other regular people.

Unfortunately, due to the scarcity of interpreters with experience translating for the deaf, hiring one is highly expensive. However, there are a few challenges that need to be addressed before a sign language recognition system can be put into place to help the deaf and hearing-impaired people. First of all, not everyone with hearing loss uses sign language for communication, which might make them feel lonely and depressed. Second, the implementation of a sign language interpreter application that could be utilized in several countries may be delayed due to the more than 200 distinct sign languages and dialects from diverse nations [4,5]. Finally, it might be overlooked that most people are not skilled in utilising cutting-edge technology because of a lack of knowledge and growth. The goal of a great deal of research and studies is supposed to be to understand as well as eliminate any barriers that individuals with auditory impairments have in order to participate in society [6]. The following examples highlight our research paper's crucial contribution:

- To improve the classification of hand motions in the American Sign Language gestures, we develop a deep learning-based method.
- To categories as well as recognise sign motions with complicated form
- discrepancies, we fitted five deep learning models. AlexNet, ConvNeXt, EfficientNet, Res- Net50, and Vision Transformer were among these models.
- We assess our scheme's performance in terms of F1-score, re-call, accuracy, and precision.
- Using the same dataset, our approach performed better than the current research.

II. LITERATURE REVIEW

Information regarding earlier studies on the use of machine learning models in the context of sign gesture and motion interpretation is gathered from the literature and presented in this section. The chapter is organised into sections that address the con- ceptual framework, relevant works, research gaps, and a summary of sign gesture interpretation. It also discusses the need of sign language and human action recognition. The related works section includes information on the alternative models that are currently in use for human action recognition and sign language recognition, along with a critical evaluation of their shortcomings and potential uses.

Nandy et al. [7] described a technique to recognise and interpret movements in Indian Sign Language from monochrome photos. Their approach [7] involves converting a film stream which includes signing motions into monochrome images, after which the characteristics are extracted using a directed spectrum. Finally, grouping is used to classify the signals into a single of several set categories according to their distinct properties. The researchers claimed a perfect sign interpretation success rate in their analysis and demonstrated that the 36-bin bitmap technique was far more precise versus the 18-bin bitmap technique.

Mekala et al. [8] introduced a deep neural network framework, which can detect and track gestures and produce writing in immediate response from a footage stream. The framework is broken down into numerous phases, such as pre-processing images, collecting data from hand movements and positions, then framing. Various attributes of a hand are indicated by its focal point of relevance [8]. The acquisition of 55 different properties employing this method served as a basis for the researchers' artificial neural network architecture, which comprised of convolutional layers which anticipated the indicators. The algorithm employed in the study was trained and tested using the whole English alphabets, starting from A to Z. They said that they had achieved 48% anti-noise immunity with a perfect recognition performance.

Chen [9] proposed an algorithm utilising an independent collection of 1200 examples of 10 unchanging signs or symbols. In the beginning, pre-processing consisted of transforming RGB images to YUQ or YIQ schemes as well as utilising boundary separation to detect boundaries as well as pigmentation [9]. The digits of a hand

which were initially detected were then identified using the convex hull method. Lastly, deep neural models were employed for the classification method. In the end, their model yielded 98.2% precision.

A method centred around Indian Sign Language was developed by Sharma et al.[10] for interacting with those with voice or auditory impairments. The true colour information had been pre-processed into monochrome employing the Matlab software once the photo was shot [10]. After that, a Sobel detection system was employed to determine the outer borders of the picture utilising a 3 × 3 filtration. Finally, an ordered clustering approach was used on the condensed image containing 600 elements, yielding 124 attributes. KNNs and deep neural systems were used as categorization techniques. The level of precision obtained with this method was 97.10%.

In order to close the communication gap between those with speech impairments and those with normal speech skills, Agarwal et al. [11] used a sensor glove for signing, processed the signs, and then presented the results in a coherent phrase. The sensor gloves were used by the individuals to execute the motions [11]. After the gestures and the database were matched, the gesture that was identified was transmitted to be processed in order to produce a sentence. The accuracy in the first iteration of the application was 33.33%. With the addition of a keyword denoting the necessary tense in version 2, 100% accuracy was attained while handling simple and continuous tens- es.

Wazalwar and Shrawankar [12] described a strategy for cropping and segmenting source footage to decode gestures. They monitored manually using the P2DHMM technique then automatically using the CamShift method. Employing a Haar Cascade algorithm, the indications were identified. Its WordNet POS tagger assigned a label to each phrase when it recognised the gestures. After that, the expression was put together through the LALR processor to create a logical English phrase.

Utilising American Sign Language (ASL), Shivashankara and Srinath [13] developed a method for recognising gestures. The study's framework, particularly used YCbCr, which increased the skin hue clustering's effectiveness [13]. This algo-rithm was used to pre-process the photos. For determining the motion, the pre- processed picture's centroid was located then the motion's maximum deviation was used to confirm it. The overall precision of this algorithm turned out to be 93.05%.

Camgoz et al. [14] published a technique showing how step-by-step visualisation, as opposed to individual mapping, is used in the interpretation of speech and gestures. The researchers introduced a new method for perception by mimicking the use of tokens and integration procedures of traditional cognitive translation systems. In the automated translation step, where it translates sign gestures to speech, the CNN architecture [14] is merged into an attention-based modulator and de-coder which simulates the odds of constructing an actual language given signing motion. The re- searchers started with the technique of word embedding, which included grouping phrases having comparable implications in order to transform a scant matrix onto a denser version. The encod-er-decoder stage was used to optimise the odds. By employing encoding, the properties associated with gestures have been generated in the form of fixed-size matrix. The phrase integrating with the earlier hidden state became the parameters used during the decoding step. This made identifying words easier. In order to prevent dependence over time and disappearing slopes, the researchers additionally included a system for concentrating throughout the decoding phase. They created an ongoing visual interpretation dataset called PHOENIX14T.

According to the study's findings, CNN could recognise sign language effectively when it used impartial training, which excluded some users and the surrounding environment. As a result, the authors suggested using CNN models to recognise sign lan- guage automatically. In a similar vein, Bhutanese sign language numerals were created using a CNN [15]. This model recognised 10 static

digits of Bhutanese sign language using around 20,000 sign pictures that were freely provided by a distinct participant. The suggested CNN model was compared with various sign languages as part of the study. Their suggested model achieved a training accuracy of 99.94 for training and 97.62% for testing based on the comparison. The number of pictures in a data set affected the misclassification and testing accuracies, according to the authors' evaluation of the model's accuracy, recall, and Fl-score. With transfer learnings like ResNet, MobileNet, and VGG16, the precision might be further adjusted.

The study mentioned above mostly shows that conduct-based methods for recognising sign language are not the best. This is due to the fact that they need comparatively costly and intricate hardware installations [16]. This explains the inclination of most researchers towards vision-based models. Despite the fact that a great deal of research has been done in this area, it is clear that developing models for the identification of sign language is still a challenging task. The hardest part is coming up with an appropriate model to solve signer-independent continuous sign issues [17, 18]. A continuous, high-accuracy model that is consistent between models is difficult to construct because of the significant variance in duration, pace, and backdrop across signers [19]. Numerous scholars have offered diverse suggestions for more investigation on models relating to CNN. Nonetheless, a further difficult task is figuring out which model to use and enhance, as several models rely on one another for the optimisation of tuning parameters [20].

III. METHODOLOGY

The four different deep learning models for recognising American alphabet motions are examined in this section. Additionally, it offers a synopsis of the dataset that our ASL recognition algorithm uses.

3.1. American Sign Language Dataset

In general, machine learning and deep learning rely substantially on data since they use algorithms to evaluate data and provide insightful predictions. One of the biggest problems confronting sign language identification and translation systems is really the scarcity of sign language databases. It is difficult to get a dataset that contains both manual and non-manual movements simultaneously. To construct and test their sign language recognition system, researchers in this subject must start from beginning with a suitably substantial database. With the aid of non-expert signers, it is simple to create a fingerspelling dataset that can be used to capture and compile sign pictures for the whole American alphabet using a standard camera. There are just 26 letters in the dictionary, and the majority of the letters are shown in a static position. We only use hand movements in our proposed system, which in sign language correspond to the American alphabet. The dataset from “IEEE Dataport” is appropriate for our approach since the finger spelling database only contains still photographs of the signer's hands. There are 87,000 photos in the IEEE Dataport collection, divided into 29 different classifications. Every class has three thousand photos in it. Twenty-six classes correspond to the twenty-six American sign language alphabets. The other classes are used for nothing, space, and deletion. The RBG format of the dataset photos comes in several forms, with size of 200 × 3200 pixels. The dataset is shown in pre- view form in Figure 1.

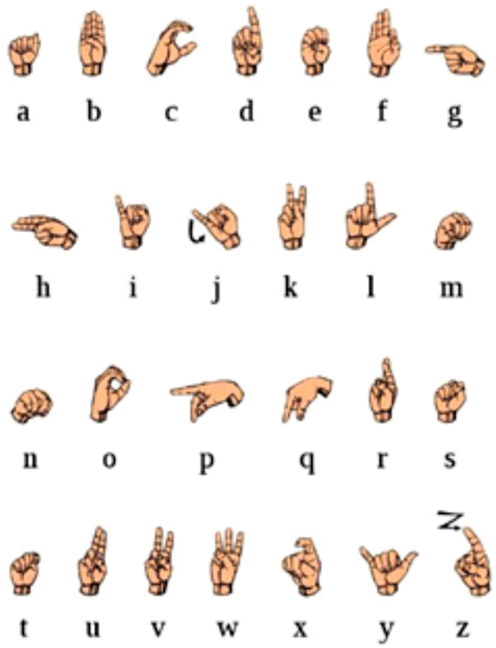

Figure 1 : American Sign Language Dataset Preview

3.2 Proposed Flow

The suggested model is shown graphically in Figure 2 for the specified application. The authors compared and determined which architecture and hyper-parameters were optimal using various GRU and LSTM. The authors built a bespoke dataset for ISL, which was used to train these models. As shown in Figure 2, the following actions were taken for every video clip.

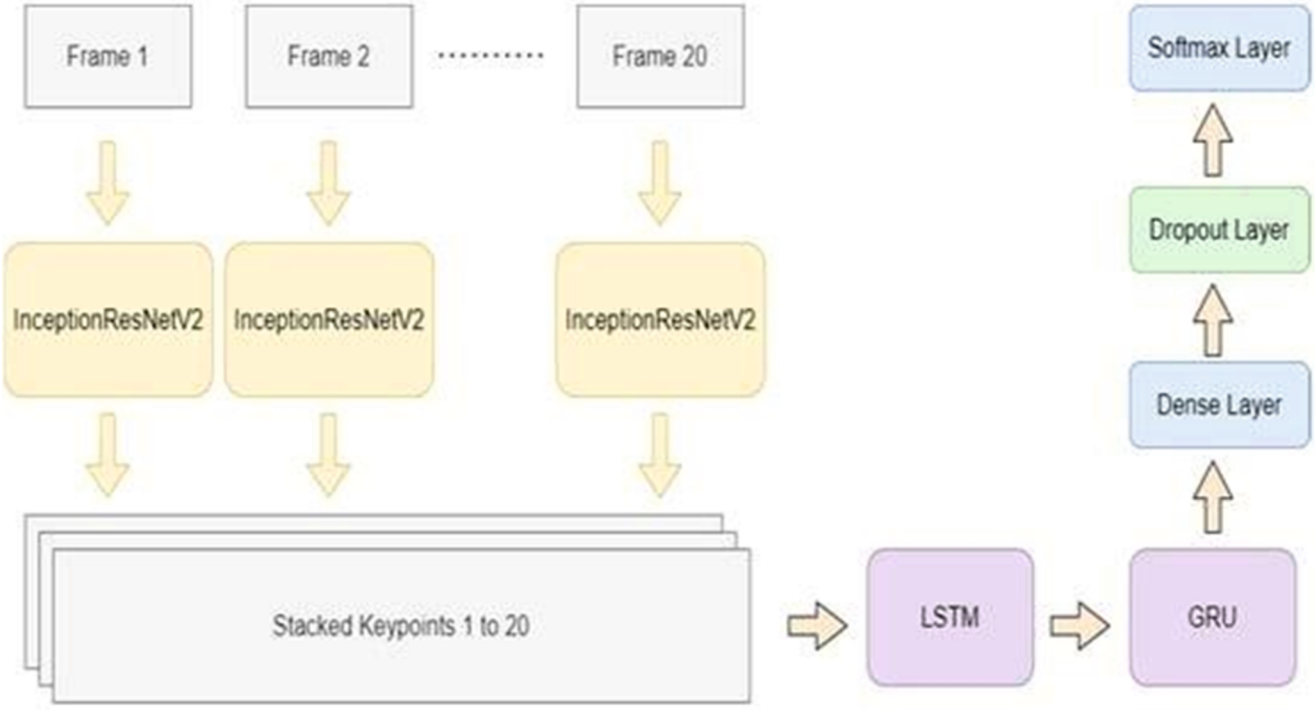

Figure 2 : Overall Architecture of the System for Gesture Recognition

1. . InceptionResNetV2 is used to extract the feature vectors, which are then sent to the model. Here, InceptionResNet-2 is used to classify the video frames into objects, and the next step is to generate key points stacked for the video frames.
2. LSTM and GRU are combined to form the neural network's first layer. This com- position may be used to more effectively capture the semantic relationships;
3. The dropout is employed to lessen overfitting and enhance the model's capacity for generalisation;
4. The "softmax" function is used to generate the final result.
5. Data from the input layer are received by the LSTM layer, which has 1536 units, 0.3 dropouts, and a kernel regularizer of 'l2'.
6. Next, the data are transmitted from the GRU layer with the same parameters;
7. A thick layer that is entirely linked receives the results;
8. The output, which has an effective value of 0.3, is sent into the dropout layer.

IV. RESULTS AND DISCUSSION

In this work, six distinct combinations of GRU and LSTM were examined: GRU then LSTM, LSTM then GRU, single-layer GRU, double-layer LSTM, and single-layer GRU. InceptionResNetV2 was first used to extract features from the movie by splitting it into frames. This produced a NumPy array value that was supplied to the training model. Subsequently, the identification of sign language was carried out. By expanding the dataset and adding more examples per word, the findings may be further enhanced and the model can be trained on more useful samples.

The Confusion matrices for the confusion matrix of the several iterations of the proposed model—two-layer GRU, two-layer LSTM, GRU-LSTM, and LSTM-

GRU—are shown in Figures 3, 4, 5, and 6. The classification model (LSTM/GRU) is evaluated using the confusion matrix. In order to determine the classification rate for input data (isolated sign language), the heat map displays the con-fusion matrix. It also has the ability to indicate mistaken identification.

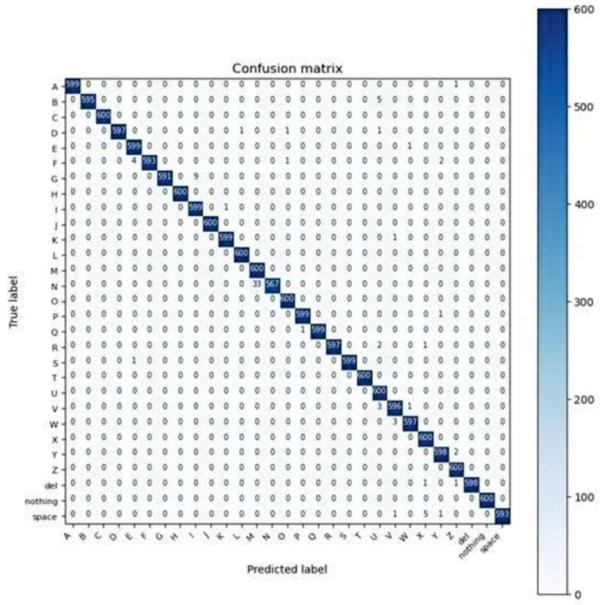

Figure 3 : Confusion Matrix of two-layer GRU

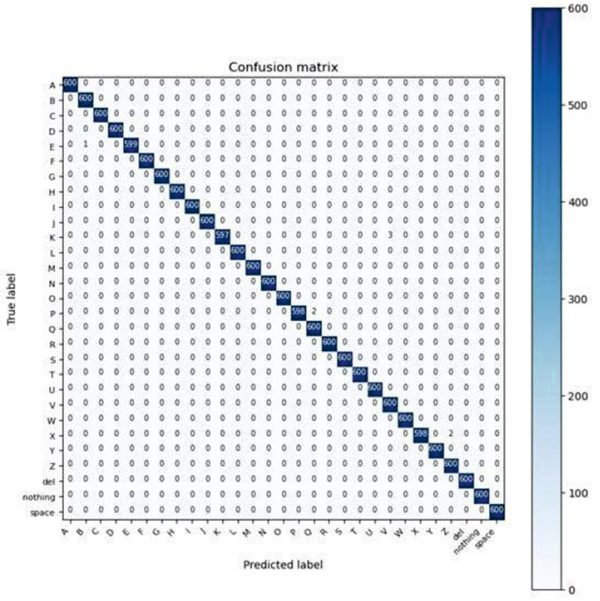

Figure 4 : Confusion Matrix of two-layer LSTM

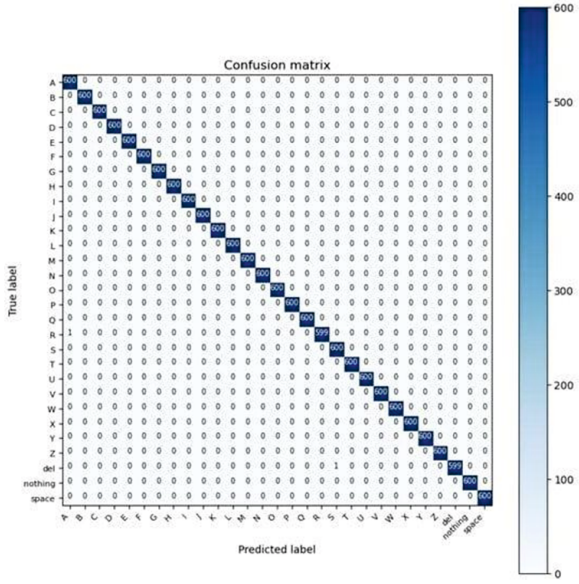

Figure 5 : Confusion Matrix of GRU followed by LSTM

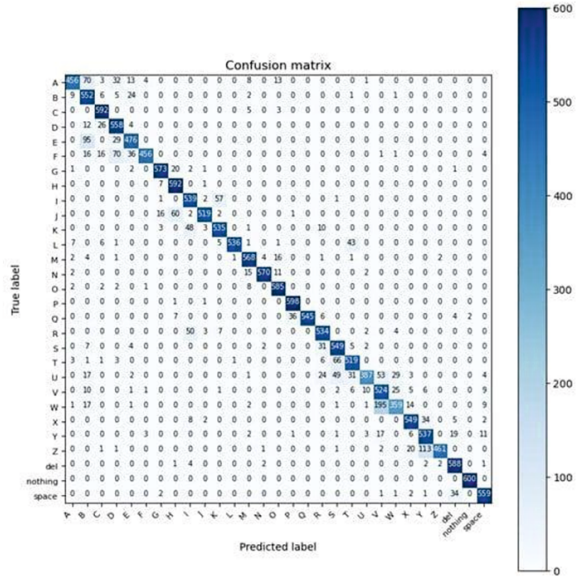

Figure 6 : Confusion Matrix of LSTM followed by GRU

The outcomes of the data model learning process are shown in Table 1, where it is evident that the suggested research model can accurately detect the signals from a variety of datasets with an accuracy rate of almost 95%. The learning and evaluation settings are the same as those used in the first configuration. (40,2048) is used as the input frame for each LSTM's 100 hidden layers, taking into account dropout, fold, learning rate, and GRU.

The outcomes of the data model learning process are shown in Table 1, where it is evident that the suggested research model can accurately detect the signals from a variety of datasets with an accuracy rate of almost 95%. The learning and evaluation settings are the same as those used in the first configuration. (40,2048) is used as the input frame for each LSTM's 100 hidden layers, taking into account dropout, fold, learning rate, and GRU.

Table 1. Performance of Various Algorithms Tested

Algorithm	L.Rate	Accuracy	Recall	F1
two-layer LSTM	0.001	97.86	98.40	97.89
two-layer GRU	0.001	98.56	98.50	98.56
LSTM-GRU	0.001	98.78	98.54	98.75
GRU-LSTM	0.001	99.04	98.70	99.01

V. CONCLUSIONS

In this study, the IEEE Dataport dataset of various hand motions is used to facilitate the identification of Indian Sign Language using LSTM and GRU. When it comes to recognising typical gestures, the suggested model works quite well. Moreover, using LSTM first and then GRU and increasing the number of layers in both models aid in improving the model's accuracy in ASL identification.

The accuracy of the model might be increased in subsequent studies by creating various datasets under optimal circumstances, shifting the camera's angle, and even using wearable technology. Although the generated models are currently limited to single signals, this method might be used to the interpretation of sign language that results in the development of syntactic elements, particularly within the framework of ASL. When vision transformers are used instead of feedback-based learning models, the outcomes may be more accurate.

REFERENCES

[1] Zwitserlood, I.; Verlinden, M.; Ros, J.; Van Der Schoot, S.; Netherlands, T. Synthetic signing for the deaf: Esign. In Proceedings of the Conference and Workshop on Assistive Technologies for Vision and Hearing Impairment, CVHI, Granada, Spain, 29 June–2 July 2004.

[2] World Health Organisation. Deafness and Hearing Loss. 2021. Available online: http://https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss (accessed on 9 November 2023).

[3] Alsaadi, Z.; Alshamani, E.; Alrehaili, M.; Alrashdi, A.A.D.; Albelwi, S.; Elfaki, A.O. A real time Arabic sign language alphabets (ArSLA) recognition model using deep learning architecture. Computers 2022, 11, 78.

[4] Alsharif, B.; Ilyas, M. Internet of Things Technologies in Healthcare for People with Hear- ing Impairments. In Proceedings of the IoT and Big Data Technologies for Health Care: Third EAI International Conference, IoTCare 2022, Virtual Event, 12–13 December 2022; Proceedings. Springer: Berlin/Heidelberg, Germany, 2023; pp. 299–308.

[5] Farooq, U.; Rahim, M.S.M.; Sabir, N.; Hussain, A.; Abid, A. Advances in machine transla- tion for sign language: Approaches, limitations, and challenges. Neural Comput. Appl. 2021, 33, 14357–14399.

[6] Latif, G.; Mohammad, N.; AlKhalaf, R.; AlKhalaf, R.; Alghazo, J.; Khan, M. An automat- ic Arabic sign language recognition system based on deep CNN: An assistive system for the deaf and hard of hearing. Int. J. Comput. Digit. Syst. 2020, 9, 715–724.

[7] Nandy, A.; Prasad, J.; Mondal, S.; Chakraborty, P.; Nandi, G. Recognition of Isolated In- dian Sign Language Gesture in Real Time. Commun. Comput. Inf. Sci. 2010, 70, 102–107.

[8] Mekala, P.; Gao, Y.; Fan, J.; Davari, A. Real-time sign language recognition based on neu- ral network architecture. In Proceedings of the IEEE 43rd Southeastern Symposium on System Theory, Auburn, AL, USA, 14–16 March 2011.

[9] Chen, J.K. Sign Language Recognition with Unsupervised Feature Learning; CS229 Pro- ject Final Report; Stanford University: Stanford, CA, USA, 2011.

[10] Sharma, M.; Pal, R.; Sahoo, A. Indian sign language recognition using neural networks and KNN classifiers. J. Eng. Appl. Sci. 2014, 9, 1255–1259.

[11] Agarwal, S.R.; Agrawal, S.B.; Latif, A.M. Article: Sentence Formation in NLP Engine on the Basis of Indian Sign Language using Hand Gestures. Int. J. Comput. Appl. 2015, 116, 18–22.

[12] Wazalwar, S.S.; Shrawankar, U. Interpretation of sign language into English using NLP techniques. J. Inf. Optim. Sci. 2017, 38, 895–910.

[13] Shivashankara, S.; Srinath, S. American Sign Language Recognition System: An Optimal Approach. Int. J. Image Graph. Signal Process. 2018, 10, 18–30.

[14] Camgoz, N.C.; Hadfield, S.; Koller, O.; Ney, H.; Bowden, R. Neural Sign Language Translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018, Salt Lake City, UT, USA, 18–22 June 2018; IEEE: Piscataway, NJ, USA, 2018.

[15] Wangchuk, K.; Riyamongkol, P.; Waranusast, R. Real-time Bhutanese Sign Language dig- its recognition system using Convolutional Neural Network. Science Direct. ICT Express 2021, 7, 215–220.

[16] Narayan, S.; Sajjan, V.S. Sign Language Recognition Using Deep Learning. In Proceed- ings of the 2021 International Conference on Intelligent Technologies (CONIT), Karna- taka, India, 25–27 June 2021; pp. 1–5.

[17] Kamal, S.M.; Chen, Y.; Li, S.; Shi, X.; Zheng, J. Technical approaches to Chinese sign language processing: A review. IEEE Access 2019, 7, 96926–96935.

[18] Gao, W.; Fang, G.; Zhao, D.; Chen, Y. A Chinese sign language recognition system based on SOFM/SRN/HMM. Pattern Recognit. 2004, 37, 2389–2402.

[19] Abdul, W.; Alsulaiman, M.; Amin, S.U.; Faisal, M.; Muhammad, G.; Albogamy, F.R.; Ghaleb, H. Intelligent real-time Arabic sign language classification using attention-based inception and BiLSTM. Comput. Electr. Eng. 2021, 95, 107395.

[20] Sharma, P.; Anand, R.S. A comprehensive evaluation of deep models and optimizers for Indian sign language recognition. Graph. Vis. Comput. 2021, 5, 200032.