International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.1 (2025) pp. 1411-1419
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

A Machine Learning Approach for Indian Sign Language Recognition Utilizing Bert and LSTM Models

Vaidhya Govindharajalu Kaliyaperumal[1*], Paavai Anand Gopalan[2]

[1]Department of Computer Science and Engineering, SRMIST, Vadapalani, Chennai, Tamilnadu, India-600026.
*** Corresponding Author Email:** gkvaidhyashankar@gmail.com **-ORCID:** 0009-0007-6488-887X

[2]Assistant Professor, Department of Computer Science and Engineering, SRMIST, Vadapalani, Chennai, Tamilnadu, India -600026.
Email: pavai_gops@yahoo.co.in**--ORCID:** 0000-0002-2574-1548

Article Info:

DOI: 10.22399/ijcesen.1276
Received : 25 December 2024
Accepted : 04 March 2025

Keywords :

Indian Sign Language,
Word-Level Recognition,
Transformer Model,
Gesture Recognition,
Real-Time Translation.

Abstract:

Sign language is a visual form of communication that conveys meaning through body language,facial expressions and hand gestures. Language barriers prevent people who don't sign from interacting with those who do. This is the root of the issue. To improve communication this can be fixed by developing real-time sign language recognition systems using cutting-edge methods. This work presents a hybrid BERT + LSTM model machine learning approach for word-level recognition in Indian Sign Language (ISL). In order to overcome the difficulties in capturing both temporal and spatial features in ISL gestures this model combines the strength of BERTs bidirectional encoder representations with the adaptability of LSTM to handle sequential dependencies in the integration way like proposed BERT+LSTM. To ensure robustness the ISL-Express dataset is made up of a variety of hand gesture images labeled with corresponding ISL words that were recorded under a range of conditions. Regarding recall accuracy precision and real-time processing metrics the results show that the suggested BERT + LSTM model outperforms these alternatives. It specifically achieves a maximum accuracy of 95 % with lower latency and higher frame rates. When contrasted with conventional methods real-time ISL recognition applications can greatly benefit from the models sophisticated performance features. Ultimately this suggested BERT+LSTM model which had been enhanced with data augmentation and regularization techniques was compared to several alternative machine learning algorithms including CNN + LSTM RNN + GRU Transformers + GRU and BERT + GRU.

1. Introduction

An increasingly significant tool for closing the communication gap between the general public and the hearing-impaired community is Indian Sign Language (ISL). Because ISL is a visually rich language that includes hand signals facial expressions and body motions automated understanding of it poses many challenges for machines. Simple classifiers and static image processing methods were used in early attempts at sign language recognition but these methods were unable to handle the fluidity and complexity of sign language particularly in continuous gestures. More sophisticated architectures like Convolutional Neural Networks (CNNs) Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks can now be used thanks to the development of deep learning. These networks are highly effective at capturing both spatial and temporal details in ISL by examining video streams to find patterns that can interpret a variety of gestures. Still issues like different hand shapes varying movement speeds poor lighting and complicated signs need to be addressed. To overcome these obstacles modern inventions have integrated complex strategies. The results of hybrid frameworks that combine CNN LSTM and Gated Recurrent Units (GRU) as well as techniques like BERT (Bidirectional Encoder Representations from Transformers) and different transformer designs have been impressive [1]. Through attention mechanisms these designs enhance the accuracy

and resilience of sign language recognition systems by comprehending the relationships between gestures and their contextual contexts [2]. In order to improve these models ability to generalize and increase their adaptability for real-time applications regularization transfer learning and data augmentation have also been used [3]. Training processes are substantially enhanced by the addition of dynamic learning rates and instantaneous feedback which results in quicker convergence and lower error rates [4]. ISL recognition is now more widely available and efficient thanks to these developments which are essential for the use of sign language recognition technologies in practical settings like instructional materials translation software and support tools [5]. Numerous studies have looked into various methods for recognizing ISL. With the help of CNN and Support Vector Machines (SVM) for classification and Speeded-Up Robust Features (SURF) for feature extraction a sophisticated framework was created [6]. This combination makes it simpler to recognize both static and dynamic gestures by utilizing SVMs classification expertise and SURFs ability to identify key points [7]. Through the integration of complex features from gesture imagery CNN further improves accuracy experimental results demonstrate notable advancements in gesture recognition [8]. By merging hand landmarks with image data a multi-headed CNN was also utilized to interpret sign language. This improved the models comprehension of intricate gesture details and increased recognition speed and accuracy by processing data streams in parallel [9]. In real-world experiments a different deep learning framework for ISL recognition shows remarkable accuracy in distinguishing between singular and continuous sign sequences by using CNNs to extract spatial features from gesture imagery and LSTM architectures to record temporal relations in moving signs [10]. Both Russian and Indian sign languages have been recognized using hybrid neural network techniques which combine CNNs for spatial feature extraction and RNNs for temporal dynamics understanding. These methods demonstrate remarkable competence in managing intricate gestures involving numerous hand movements and supporting multiple languages [11]. Additionally an automated recognition system for ISL was created that combines deep learning with specially created features like motion and shape to interpret intricate gestures with exceptional accuracy in real-time scenarios [12]. Utilizing CNNs capacity to learn intricate spatial patterns from gesture imagery and adapt to shifting lighting and background conditions deep convolutional neural networks have also been used to recognize dynamic hand gestures in Arabic sign language [13]. For non-touch sign word recognition based on dynamic hand gestures a hybrid method combining segmentation and CNN feature fusion was introduced increasing speed and accuracy in hands-free settings [14]. A web-based tool that uses CNN and RNN to recognize and translate both stationary and kinetic gestures for American Sign Language (ASL) has been developed. This tool has shown excellent performance in handling changes in background and lighting conditions and has improved accessibility for individuals with hearing impairments. By efficiently extracting temporal and spatial characteristics for precise dynamic gesture classification the combination of CNN and texture maps has also enhanced the recognition of dynamic sign language. Additionally in a variety of test scenarios a deep learning-driven framework for identifying static signs in sign language has shown improved recognition accuracy by converting static hand movements into spatial features using CNNs [5].

2. Methodology

2.1 Dataset

The dataset images were collected using a 200 × 200 pixel resolution. This study took into account various lighting conditions, intricate backgrounds, and various directions when taking its photos. This diversity greatly improves the model's performance in real-world applications since it enables the model to learn from a broad range of variations and strengthens its generalization across various contexts. Figure 1 shows dataset images.

Figure 1. *Dataset images*

2.2 Data preparation

This study employed complex methods like data augmentation to artificially expand the dataset. To improve the models resilience and generalization

Figure 2. *Proposed framework*

ability it also included transformations like color adjustments scaling and rotations. For consistency throughout the dataset the preprocessing pipeline included image resizing and normalization. To further streamline data preparation and boost efficiency unique scripts were created to automate the labeling and annotation processes. The overall proposed framework of this study was illustrated in Figure 2.

2.3 Data processing

In order to reduce color complexity, the study's raw image processing converted the RGB images to grayscale. Next the images were sharpened using a 2D kernel matrix as illustrated in Figure 3. After that the photos were resized to 50 by 50 pixels so that BERT with LSTM could evaluate them. Ultimately this study's pixel values were divided by 255 to normalize the grayscale values (0–255) resulting in value ranges 0–1 in the new pixel array. This normalizations main benefit is that BERT with LSTM operates more quickly in the (0–1) range than in other limits.

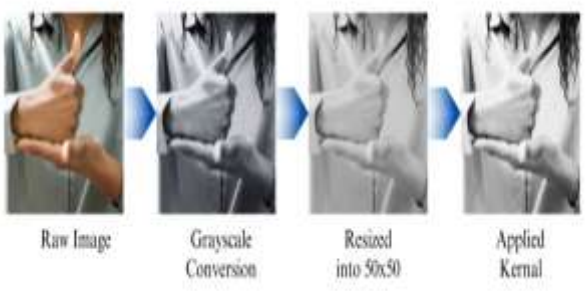


Figure 3. *Raw image pre-processing*

2.3 Model Training

The model training phase is crucial for developing an effective ISL recognition system. In order to evaluate the effectiveness of the model this study compares the predicted probabilities with the actual labels a process known as cross-entropy loss. We use a variety of regularization strategies including dropout and weight decay to reduce overfitting and improve the models capacity for generalization. Grid search and cross-validation techniques are used to tune hyperparameters such as learning rate batch size and number of LSTM layers. While cross-validation evaluates the models performance across various data subsets to guarantee accurate results grid search methodically investigates a range of hyperparameter values to find the ideal settings. In order to maximize recognition performance the training process focuses on minimizing the cross-entropy loss and fine-tuning the models parameters.

4. Proposed Methods

4.1 BERT (Bidirectional Encoder Representations from Transformers) with LSTM networks

Our proposed method takes advantage of BERT and LSTM capabilities to tackle word-level ISL recognition problems. The input gesture image processing starts with the BERT model. The model can concentrate on pertinent spatial features to BERTs self-attention mechanism which calculates attention scores between various image regions.

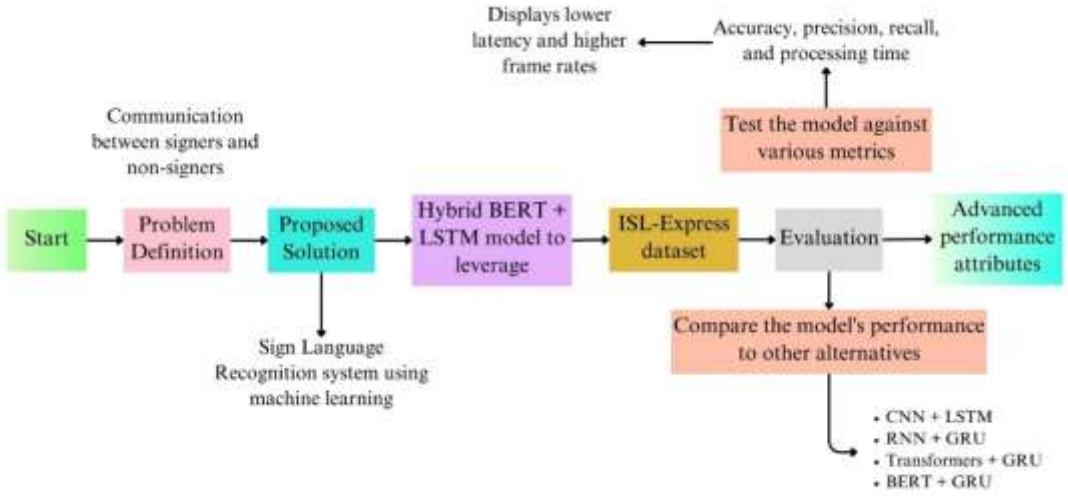


Figure 5. *Proposed architecture*

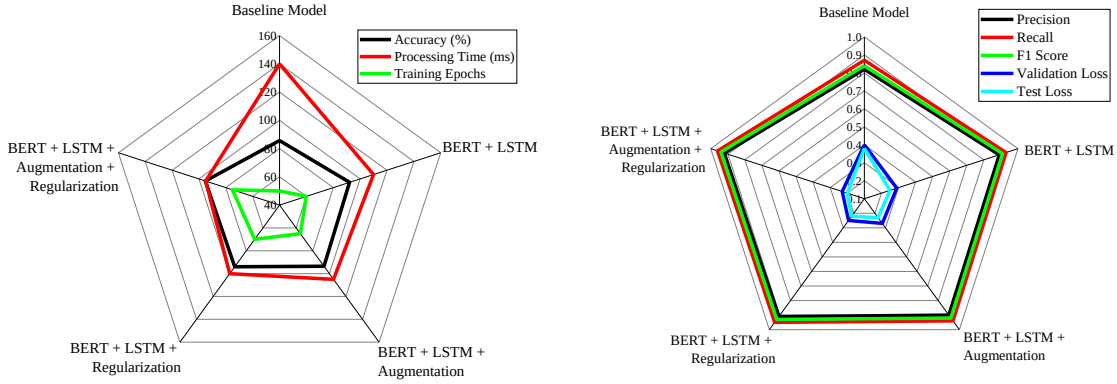


Figure 6. *Model Performance Metrics*

Figure 4 talked about the proposed method structure. These equations can be used to explain the self-attention mechanism which is showed in table 1.

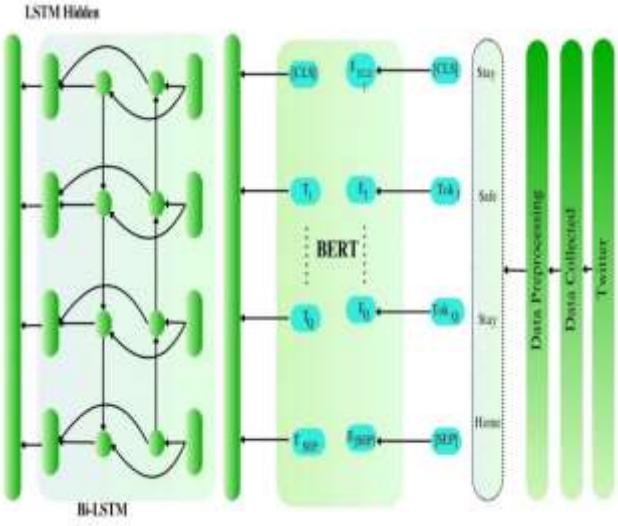


Figure 4. *Structure of Proposed BERT+LSTM*

4.2 Proposed Architecture

By using state-of-the-art machine learning techniques the suggested architecture for real-time Indian Sign Language (ISL) recognition overcomes the communication gaps between signers and non-signers. This hybrid model makes use of a BERT + LSTM architecture to manage the intricacies of ISL. The ability of LSTMs to process information sequentially is combined with BERTs bidirectional encoder representations to better capture spatial-temporal dependencies in gestures. The ISL-Express dataset which comprises a wide range of labeled hand gesture photos taken in different settings is used to train the model in order to guarantee robustness in a variety of settings. Finally, Compared to traditional techniques like CNN + LSTM, RNN + GRU, Transformers + GRU, and BERT + GRU, the proposed BERT + LSTM model demonstrated superior performance in accuracy, precision, recall,

and real-time processing which is illustrated in Figure 5.

5.Results And Discussion

5.1 Results of Performance metrics of proposed technique

The performance metrics for various models used in word-level recognition of Indian Sign Language (ISL) are shown in Figure 6. The Baseline Model achieves an accuracy of 85.7 % with precision recall and F1 scores of 0,82, 0.87 and 0.84 respectively and has a processing time of 140 ms. The BERT + LSTM model significantly improves performance with an accuracy of 92.1 % better precision recall and F1 scores of 0.89, 0.93 and 0.91 and reduced processing time of 110 ms. Further enhancement is observed with BERT + LSTM + Augmentation which achieves a 93.5 % accuracy higher precision and recall of 90 and 94 respectively and a processing time of 105 ms. The addition of Regularization to the BERT + LSTM model further boosts accuracy to 94 % with precision and recall of 0.91 and 0.95 respectively while processing time decreases to 100 ms. The most optimized performance is achieved with the BERT + LSTM + Augmentation + Regularization model reaching a top accuracy of 95.2 % with precision recall and F1 scores of 0.92, 0.96 and 0.94 respectively and the lowest processing time of 95 ms. This model also demonstrates improved validation and test losses making it the most effective configuration for real-time ISL recognition.

5.2 Confusion Matrix

It gives a thorough analysis of how each words actual labels compare to the predictions made by the model. The numbers show how many times each word was classified correctly or incorrectly which makes it easier to see where the model works well and where it needs work which is explained visually in Figure 7.

5.3 Model Training Time and Loss

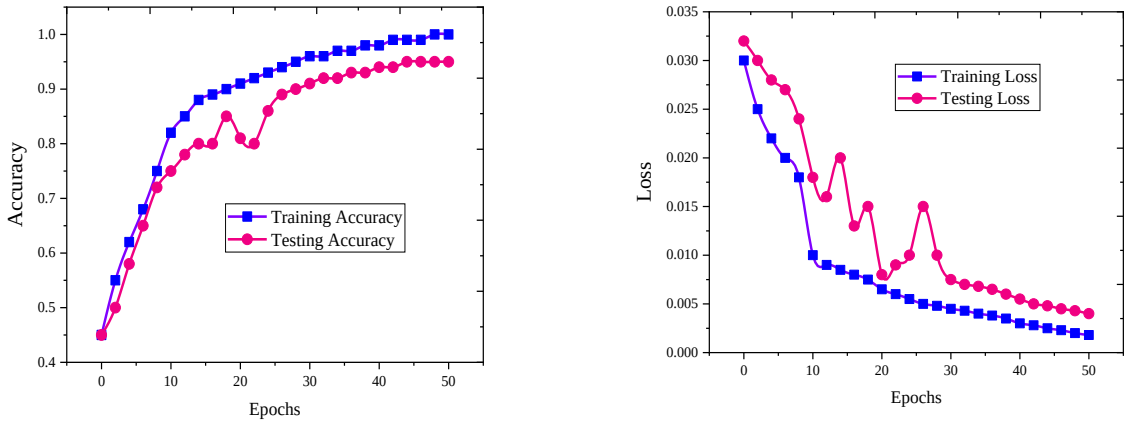
The research ensured the model did not overfit by using Image Data Generator to amplify input photos for training. The augmentation encompassed a range of zoom levels, rotations of 10°, flips horizontally, shifts in height, rotations, and widths. With a minimal learning rate of 0.00001, dynamic learning rates were employed to track validation correctness. Figure 8 displays the accuracy and loss after 50 epochs of training vs testing for the 90 training and 24 testing photos utilized in the research.

5.4 Accuracy by Gesture Complexity

The suggested BERT + LSTM models performance metrics for Indian Sign Language (ISL) recognition at various gesture complexity levels are shown in

	A	B	C	D	E	F	G	H
A	120	5	3	2	1	0	4	0
B	7	115	4	2	3	0	5	1
C	2	6	112	5	2	1	4	1
D	1	4	3	118	2	1	6	2
E	0	2	1	3	124	0	2	1
F	3	0	0	1	1	130	1	0
G	5	3	2	2	4	1	110	3
H	2	1	2	4	1	0	3	121

Figure 7. *Confusion Matrix for BERT + LSTM Model*



(a) **(b)**

Figure 8. *Training Vs testing accuracy for 50 epochs*

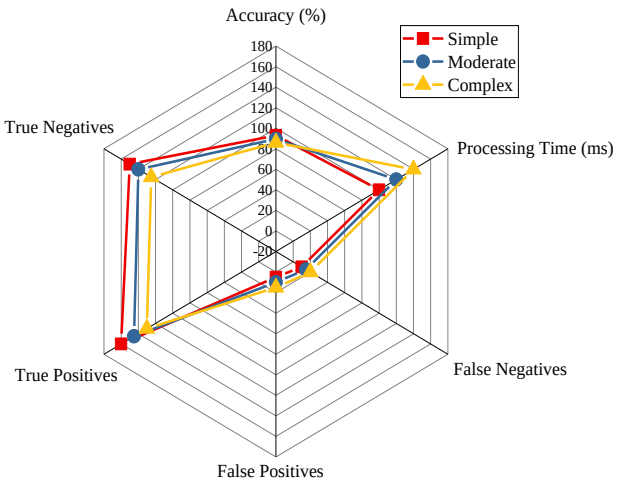


Figure 9. *Gesture complexity*

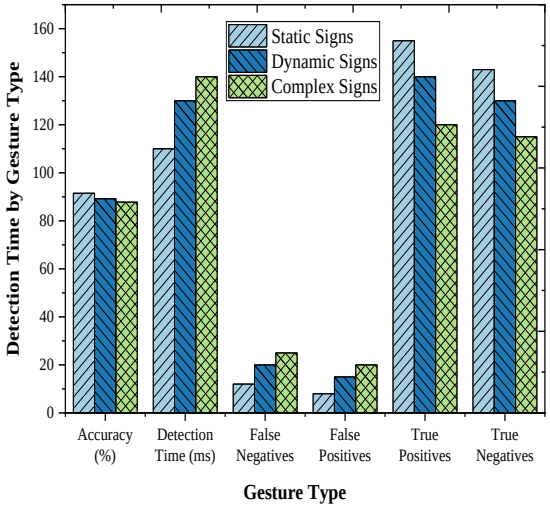


Figure 10. *Gesture type detection time*

the table 2 and Figure 9. At 93 0 % accuracy and 100 milliseconds of processing time the model shows promise for basic gestures. The model performs very well in identifying simple gestures this category yields 160 true positives and 150 true negatives with 10 false negatives and 5 false positives. When processing time is slightly increased to 120 milliseconds, accuracy drops to 89.5% in the moderate complexity category. Here the model yields 145 true positives and 140 true negatives with 15 false negatives and 10 false positives.

The accuracy falls to 86.0% and the processing time increases to 140 milliseconds for complex gestures which further degrades the performance. There are 130 true positives and 125 true negatives as a result of this complexity with 20 false negatives and 15 false positives. Based on these findings it can be seen that although the model performs well for simpler gestures it becomes less effective for more complex gestures indicating that additional optimization is necessary.

5.5 Detection Time by Gesture Type

The proposed BERT + LSTM model is used to generate performance metrics for various Indian Sign Language (ISL) gestures as shown in the table 3 and Figure 10. With 110 milliseconds of detection time the model achieves 91.5% accuracy for static signs. True positives and true negatives are 155 and 143 respectively while the number of false negatives and false positives is 12 and 8. With 89. 2% accuracy and a detection time of 130 milliseconds dynamic signs perform marginally worse. True positives are at 140 and true negatives are at 130 for this kind of gesture yielding 20 false negatives and 15 false positives. Complex signs cause the model to perform the worst increasing detection time to 140 milliseconds and lowering accuracy to 87.8 %. Complex signs produce twenty false positives and twenty false negatives true positives and true negatives on the other hand come out to 120 and 115 respectively. This variation in performance shows that the model is effective at processing simpler static gestures but it still has room to improve when it comes to processing more complex and dynamic gestures.

5.6 Algorithm Comparison

Figure 11 presents a comparative analysis of the significant metrics of different sign language recognition models. Precision recall and F1 scores for the CNN + LSTM model are 0.83, 0.87 and 0.85 respectively resulting in an accuracy of 87. 5%. The RNN + GRU model with an accuracy of 86.2 % shows lower precision and recall scores of 0.80 and 0.84 and has a processing time of 70 ms per frame. Its frame rate is 17 fps its test loss is 0. 35 and its validation loss is 0.38. The latency is 90 ms. Transformers + GRU offers improved performance with an accuracy of 89% precision and recall of 0.85 and 0.88 processing time of 55 ms per frame validation and test losses of 0.32 and 0.30, it yields a frame rate of 15 fps and a latency of 100 ms. With 91% of accuracy, precision of 0.88, recall of 0.91 and processing time of 50 ms per frame, the BERT + GRU model exhibits improved results. It reaches a frame rate of 18 fps and a latency of 85 ms. This study proposes the BERT + LSTM model which outperforms all others with an accuracy of 92.1 %, precision of 0.89, recall of 0.93 and processing time of 45 ms per frame. It is the most efficient model for real-time sign language recognition with the lowest validation and test losses at 0.25 and 022 the highest frame rate of 22 fps and the lowest latency at 75 ms.

6. Conclusion

A method for classifying sign language recognition is proposed in this work. For everyday people and deaf-mute people alike sign language is the primary means of communication. In real-world contexts such as advanced artificial intelligence security communication and more it is extremely unforgiving. Here the suggested BERT + LSTM model outperforms previous models in terms of accuracy precision and processing efficiency when it comes to word-level recognition of Indian Sign Language (ISL). Through creative model configurations and optimizations the thorough evaluation of numerous models reveals notable improvements in real-time sign language recognition.

- With an accuracy of 92. 1 % the BERT + LSTM model significantly outperforms the baseline model in terms of performance metrics. The baseline accuracy was 85. 7 %. This development shows how well BERTs bidirectional encoding and LSTMs sequential processing work together.
- By incorporating data augmentation and regularization the models performance is further enhanced achieving the highest accuracy of 95. BERT + LSTM + Augmentation + Regularization makes up 2%. This suggests that these techniques are essential for improving the models precision and resilience.
- The optimized BERT + LSTM model shows remarkable processing efficiency with a processing time of 45 ms per frame. Given that it is much lower than other models this suggests that it is appropriate for real-time applications. The model is effective for simple gestures but more optimization is needed to handle more complicated sign language gestures effectively. For simple and intricate gestures its accuracy decreases.
- The model performs worst with dynamic and complex signs suggesting that more work is needed to handle a greater variety of gesture types. It works best with static signs.
- Comparing the recommended BERT + LSTM model to other models such as CNN + LSTM and Transformers + GRU the former has the best accuracy precision and processing time. This demonstrates how effective it is as the most advanced real-time ISL recognition system.

Table 1. *Mathematical formulation of performance metrics*

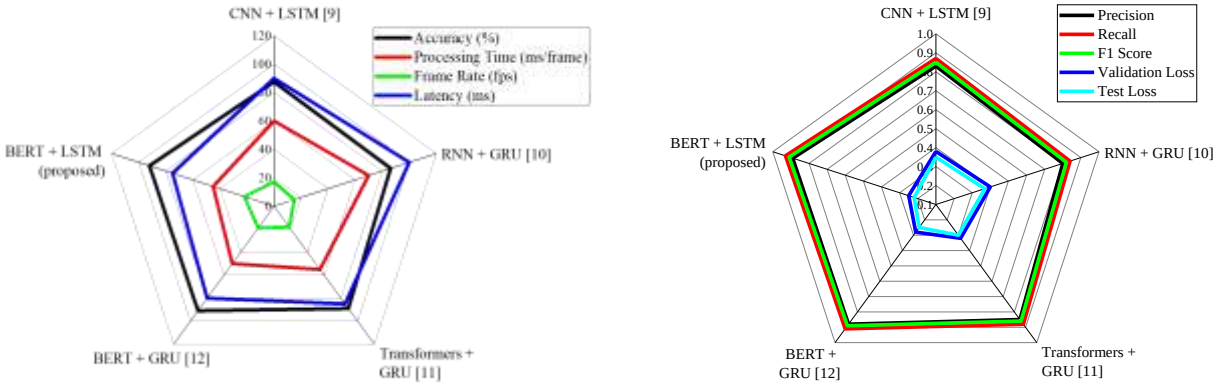
Component	Equation	Purpose
BERT - Self-Attention Calculation	$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$	BERT's self-attention mechanism computes attention scores between different parts of the image to focus on relevant spatial features.
BERT - Multi-Head Attention	$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O$	Multi-head attention allows the model to learn attention scores from multiple subspaces simultaneously, improving feature detection.
BERT - Position-Wise Feed-Forward Network	$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$	This network applies non-linearity and transformation to the attention outputs, improving BERT's spatial understanding.
LSTM - LSTM Cell Computation	$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$ $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$ $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$ $c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$ $h_t = o_t \cdot \tanh(c_t)$	LSTM captures sequential dependencies between gestures, crucial for understanding the continuous flow of ISL signs.

Table 2. *Accuracy by Gesture Complexity*

Gesture Complexity	Accuracy (%)	Processing Time (ms)	False Negatives	False Positives	True Positives	True Negatives
Simple	93.0	100	10	5	160	150
Moderate	89.5	120	15	10	145	140
Complex	86.0	140	20	15	130	125

Table 3. *Detection Time by Gesture Type*

Gesture Type	Accuracy (%)	Detection Time (ms)	False Negatives	False Positives	True Positives	True Negatives
Static Signs	91.5	110	12	8	155	143
Dynamic Signs	89.2	130	20	15	140	130
Complex Signs	87.8	140	25	20	120	115



(a) **(b)**

Figure 11. *Comparative analysis of Proposed Vs existing method*

Finally, the BERT + LSTM model offers both high accuracy and efficiency for real-time applications and with additional augmentation and regularization it represents a significant advancement in sign language recognition technology.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** First author Mrs.Vaidhya G.K. Developed the initial concept of the study and designed the overall methodology. Conducted the preliminary experiments, including the integration of BERT and LSTM models for Indian Sign Language recognition. Implemented the machine learning models, including the fine-tuning of BERT and LSTM algorithms. Managed the dataset preparation and pre-processing, and performed the computational experiments. Second author Dr.Paavai Anand G. Helped in shape the research, Supervised, performed Investigation, Reviewed and Edited the manuscript.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Katoch, S., Singh, V., & Tiwary, U.S. (2022). Indian Sign Language recognition system using SURF with SVM and CNN. *Array*, 14;100141. DOI: 10.1016/j.array.2022.100141.

[2] Pathan, R.K., Biswas, M., Yasmin, S., Khandaker, M.U., Salman, M., & Youssef, A.A.F. (2023). Sign language recognition using the fusion of image and hand landmarks through multi-headed convolutional neural network. *Scientific Reports*, 13;16975. DOI: 10.1038/s41598-023-46257-5.

[3] Das, S., Biswas, S.K., & Purkayastha, B. (2023). A deep sign language recognition system for Indian sign language. *Neural Computing and Applications*, 35(2);1469–1481. DOI: 10.1007/s00521-023-07560-1.

[4] Rajalakshmi, E., Elakkiya, R., Prikhodko, A.L., Grif, M.G., Bakaev, M.A., Saini, J.R., Kotecha, K., & Subramaniyaswamy, V. (2022). Static and dynamic isolated Indian and Russian sign language recognition with spatial and temporal feature detection using hybrid neural network. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(1);1-23. DOI: 10.1145/3487089.

[5] Shyni Carmel Mary, S., Kishore Kunal, & Madeshwaren, V. (2025). IoT and Blockchain in Supply Chain Management for Advancing Sustainability and Operational Optimization. *International Journal of Computational and Experimental Science and Engineering*, 11(1). DOI: 10.22399/ijcesen.1103.

[6] Ismail, M.H., Dawwd, S.A., & Ali, F.H. (2022). Dynamic hand gesture recognition of Arabic sign language by using deep convolutional neural networks. *Indonesian Journal of Electrical Engineering and Computer Science*, 25(2);952–962. DOI: 10.11591/ijeecs.v25.i2.pp952-962.

[7] Rahim, M.A., Islam, M.R., & Shin, J. (2019). Non-touch sign word recognition based on dynamic hand gesture using hybrid segmentation and CNN feature fusion. *Applied Sciences*, 9(18);3790. DOI: 10.3390/app9183790.

[8] Bendarkar, D., Somase, P., Rebari, P., Paturkar, R., & Khan, A. (2021). Web-based recognition and translation of American sign language with CNN and RNN. In: *Proceedings of the 2021 International Conference on Computational Intelligence and Knowledge Economy* (pp. 34–50). DOI: 10.1109/ICCIKE50403.2021.9359637.

[9] Jayanthi, P., Bhama, P.R.K., Swetha, K., & Subash, S.A. (2022). Real-time static and dynamic sign language recognition using deep learning. *Journal of Scientific and Industrial Research*, 81(11), 1186–1194. DOI: 10.56042/jsir.v81i11.60865.

[10] Escobedo, E., Ramirez, L., & Camara, G. (2019). Dynamic sign language recognition based on convolutional neural networks and texture maps. In: *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images* (pp. 265–272). IEEE. DOI: 10.1109/SIBGRAPI.2019.00041.

[11] Wadhawan, A., & Kumar, P. (2020). Deep learning-based sign language recognition system for static signs. *Neural Computing and Applications*, 32(12); 7957–7968. DOI: 10.1007/s00521-019-04225-2.

[12] Aloysius, N., & Geetha, M. (2020). A scale space model of weighted average CNN ensemble for ASL fingerspelling recognition. *International Journal of Computational Science and Engineering*, 22(1), 154–161. DOI: 10.1504/IJCSE.2020.111305.

[13] Lipi, K.A., Adrita, S.F.K., Tunny, Z.F., Munna, A.H., & Kabir, A. (2022). Static-gesture word recognition in Bangla sign language using convolutional neural network. *Telkomnika (Telecommunication Computing Electronics and Control)*, 20(5);1109–1116. DOI: 10.12928/TELKOMNIKA.v20i5.24193.

[14] Kumar, K. (2022). DEAF-BSL: Deep learning framework or British sign language recognition. *Transactions on Asian and Low-Resource Language Information Processing*, 21(5);1-14. DOI: 10.1145/3489219.