Foundation University Journal of Engineering and Applied Sciences

FUJEAS
Vol. 6, Issue 1, 2025
DOI:10.33897/fujeas.v6i1.880

Research Article

Comparative Analysis of GRU and LSTM-based Models for Pose Estimation in Pakistan Sign Language Recognition

Safa Khan [a], Akbar Hussain[*, b], Ishal Imran [a], Hirra Shahbaz [a], Rafia Amjad [a], Mujeeb Ur Rehman [b]

[a] Department of Artificial Intelligence, University of Management & Technology (UMT), Sialkot, Pakistan
[b] Department of Information Technology, University of Management & Technology (UMT), Sialkot, Pakistan

[*] **Corresponding author**: akbar.hussain@skt.umt.edu.pk

Article Citation:
Hussain et al. (2025). "Comparative Analysis of GRU and LSTM based Models for Pose Estimation in Pakistan Sign Language Recognition". *Foundation University Journal of Engineering and Applied Sciences*
DOI:10.33897/fujeas.v6i1.880

Published by
Foundation University
Islamabad.
Web: https://fui.edu.pk/

Abstract:

This study explores Sign Language Recognition (SLR) within the context of Pakistan Sign Language (PSL), aiming to bridge communication gaps between signers and non-signers. Sign languages employ handshapes, body gestures, and facial expressions to facilitate communication, addressing the worldwide linguistic needs of deaf communities. While significant efforts have been devoted to global SLR and Sign Language Translation (SLT) systems, limited attention has been paid to PSL. To address this gap, we propose a novel approach for dynamic word-level SLR, incorporating manual and non-manual features. The proposed method utilizes pose estimation RNN-based architectures (GRU and LSTM) on both our proprietary pronoun-based video dataset and the PkSLMNM dataset. By extracting key points from 3D coordinates within individuals, we propose several optimization functions for original and augmented datasets. We then compare the sequential classification potential of GRUs and LSTMs. Our findings reveal that GRU outperforms LSTM, achieving a 4% improvement in real-time classification accuracy on both augmented and original datasets, with an overall accuracy of 98.61%.

Keywords: LSTM; Pakistan Sign Language; SLR; RNN; Sign Language Translation; Urdu Language.

1. Introduction

Sign language is a language that helps deaf individuals who are unable to speak communicate through gestures. These gestures are made through handshapes, body movements, and facial expressions. Just like spoken languages, sign languages differ by country and are specific to their regions. The estimation suggests that there are over 72 million people [1] with hearing disabilities, and out of them, 10 million are from Pakistan [2]. There are a total of 300 sign languages in the world. Sign languages such as American Sign Language (ASL), British Sign Language (BSL), Arabic Sign Language (ArSL), German Sign Language (DGS), Chinese Sign Language (CSL), Pakistan Sign Language (PSL), and many more contribute to each country's deaf society. For example, Pakistan Sign Language (PSL) is the primary SL of Pakistan, but its usage and regional dialects vary for each province, such as for Sindh, Punjab, Balochistan, and Khyber Pakhtunkhwa (KPK), each of which has its own PSL. Researchers have made efforts to bridge communication barriers between signers and non-signers by introducing Sign Language Recognition (SLR) and Sign Language Translation (SLT) technology. SLR is the procedure of automatically interpreting and understanding SL gestures performed by signers. It pertains to developing algorithms that can analyze video, image, or signal data of captured SL gestures and translate them into spoken language text, speech, symbols, or visually represent gestures in the form of animated avatars. Notably, SLR experienced interest in the early 1990s [3]. Whereas SLT involves converting recognized signs into spoken language text, although it can be carried out independently, such as Sign2Text (S2T) [4].

There are two widely used approaches for SLR: Vision-based and Sensing-based. Vision-based approach requires two forms of input, such as image and video data. Image input is preferably used for signs which are static, machine learning and neural networks based supervised learning with the use of Hidden Markov Model (HMM), Support Vector Machine (SVM), Convolutional Neural Network (CNN), 2D-CNNs K-Nearest Neighbour (KNN) and random forest based classifications are greatly utilized where model learns features of raw pixel data for Isolated Sign Language Recognition (ISLR). Video input is required when gestures are dynamic and are classified through seizing temporal dynamics and the sequential nature of signs with 3D-CNNs, Recurrent Neural Networks (RNN), Long-Short Term Memory (LSTM), Temporal Convolutional Network (TCN), and Transformers, progressing the Continuous Sign Language Recognition (CSLR) research.

A sensing-based approach takes gesture input through sensors. Researchers may employ gloves like Datagloves or Cybergloves, which are embedded with sensors, while others may use a customized number of sensors to detect finger flexes, movement, and the location of gestures. A fundamental unit of SL is known as a gloss, which serves as a mimicry of a sign into spoken language [5]. As compared to other SLs like DGS, ASL [6], and CSL [7], prominent work has been done on ISLR, including word or fingerspelling-level recognition and CSLR involving a sentence-level recognition being the engrossment of researchers due to an extensive amount of datasets available, i.e., PHOENIX-2014T [8], SIGNUM [9] and CSL-Daily [10] including gloss and translation annotations, whereas on PSL there is an expand of researches on fingerspelling albeit a contemporary regard on word-level but a lack of focus on sentence-level recognition. This is primarily due to the availability of only fingerspelling datasets. A recent contribution of the PKSMLN dataset [11] introduced a word-level video dataset incorporating both manual and non-manual features, despite encountering some inconsistencies in certain frames. Nevertheless, this dataset represents a notable advancement as the first large-scale video dataset for PSL. The work is limited with respect to PSL due to the requirement of publicly available datasets, which is so important to improve the field of SLR and SLT on PSL.

In this paper, a sign language recognition system is proposed for PSL, which first contributes to the SLR by developing the first dataset of seven word-level pronouns incorporating both manual and non-manual features without a constricted background. We also included the PKSMLNM dataset [11] to improve our training data's potency. Mediapipe Holistic is implemented to extract hand, body, and facial landmarks from a holistic view of both datasets. LSTM and GRU models, due to their ability to capture temporal dependencies in sequential data, are trained with preprocessed features, allowing us to compare their accuracies and performance. This study generally provides an insight into the model's capability of handling large datasets for dynamic analysis of gestures with varying backgrounds and limited computational resources for real-time sign language recognition.

The proposed study aims to pioneer SLR in Pakistani Sign Language (PSL) by leveraging pose estimation techniques and novel pronoun datasets. We work on extracting key point features from RGB video data for real-time recognition, optimizing LSTM and GRU model architectures for efficient gesture classification. The goal of the research is to evaluate the effectiveness of Mediapipe's Holistic pose estimation, analyzing critical parameters for accurate recognition, and evaluating model performance on both datasets. Furthermore, we aim to identify computationally efficient real-time recognition methods and conduct a comparative analysis to determine the most effective sign language recognition strategies. The key contributions of the study are listed below.

- Creation of the first dataset consisting of seven word-level pronouns of Pakistani Sign

Language (PSL), addressing a significant gap in available resources for PSL recognition.

- The data set combines manual and non-manual features without background restrictions, increasing the richness and diversity of PSL recognition training data.
- The integration of the PKSMLNM dataset increases the potential of training data and helps improve model performance and accuracy of sign language recognition.
- Use Mediapipe Holistic to implement pose estimation, which can extract signatures, bodies, and faces from a holistic perspective, thereby improving the feature extraction process for both datasets.
- This study evaluates the ability of LSTM and GRU models to capture temporal dependencies in sequential data, providing valuable insights into their accuracy and performance by optimizing models in real-time sign language recognition.

The rest of the paper is organized as follows: Section 2 discusses the related work of the paper, Section 3 discusses the proposed methodology, Section 4 presents results and discussion, and finally Section 5 concludes the work.

2. Related Work

The authors in [13] introduced their own dataset comprising 6633 images of thirty-six single-handed static alphabets, developed by six signers. They used feature extraction methods such as HOG, EOH, LBP, and SURF, and then compared them using the Multiple Kernel Learning (MKL) technique with linear, polynomial, and Gaussian kernel functions in SVM classification. The highest accuracy, 89.52%, was achieved with HOG using a linear kernel function. Despite successful classification results, their dataset has limitations, including constraints on background appearance, clothing, limited distance, and involvement of only static alphabets with a single hand.

Another study [14] also proposed their dataset comprising thirty-seven Urdu alphabets. The images were annotated using a classification system, specifically SVM, and stored in XML file format. For detection, they utilized shape classification methods including Fast Fourier Transform (FFT) for rotation invariance and energy normalization for scale invariance. Recognition was achieved through a one-against-all strategy, with SVM training and testing based on the shape of each hand configuration's signature, with 80-90% accuracy, despite the research being limited to single-handed and static gesture recognition.

Another dataset was generated for static Urdu numbers utilizing both palmar and dorsal sides. Preprocessing steps were applied to minimize noise, including the addition of an additional picture in the background and its removal from the original image. A Bag-of-Words (BoW) technique was used to construct histograms for feature extraction by [15]. Evaluation using Random Forest, SVM, and KNN classifiers achieved accuracies of 88%, 90%, and 84%.

Another study [16] also proposed a PSL recognition system using BoW and SVM techniques. They curated a dataset involving 36 static and 3 dynamic Urdu alphabets, but used Speeded Up Robust Feature (SURF) descriptors and BoW representation for feature extraction. Their system achieved accuracies of 97.80% for static signs and 96.53% for dynamic signs. However, they imposed restrictions on the background and clothing color, limiting generalization and potentially introducing bias towards specific hand shapes or movements.

In this study, [17] a pipeline for the recognition of PSL is introduced, integrating an augmentation unit covering adjustments in brightness, contrast, noise, rotation, scaling, and translation. They made use of the PSL dictionary dataset, consisting of 80 commonly used signed words, each with two samples. To assess the efficacy of their proposed pipeline, three deep learning models—C3D, I3D, and TSM—were proposed. Findings indicate that translation and rotation are the most effective augmentation techniques. Models trained using their data-augmented pipeline outperformed other methods relying solely on original data. The C3D model exhibited the highest suitability, achieving an accuracy of 93.33% while requiring less training time compared to other models.

This paper [18] also introduced their own dataset, captured in real-time through webcam, consisting of static numbers 1-10, as well as "OK" and "Salaam" gestures, comprising nine distinct hand movements with 500 images for each gesture. Through image preprocessing techniques, including greyscale conversion, ROI determination, background subtraction, and contour analysis, the system predicts gestures using CNN and achieves an average accuracy of 98.76% in real-time hand gesture identification.

According to [19], a method of an end-to-end SLR utilizing LSTM for CSL was proposed. Their system processes the moving trajectories of 4 skeleton joints, eliminating the need for explicit feature design. Evaluation on a large isolated CSL vocabulary dataset captured by Kinect 2.0 demonstrates the superiority of their approach over HMM methods. Another study [20] conducted on Spanish Sign Language recognition into text used LSTM to address the challenge of recognizing non-static signs through deep learning, particularly focusing on action detection by analyzing hand, face, and pose cues. The system was trained on a dataset comprising 330 videos, achieving an impressive accuracy of 98.8% across five sign classes. Authors of [21] proposed dynamic gesture recognition using 3DCNN+ConvLSTM and achieved high accuracy, thereby reducing training time significantly.

Similar work of [22] used Mediapipe Holistic in recognizing multiple datasets such as ASL, ISL, and Italian Sign Language, through real-time detection using SVM with a higher accuracy of 99% to other deep learning models such as ANN and MLP. This study [23] proposed the MOGRU method, involving MediaPipe and a GRU model, for Indian sign language recognition, and optimized a standard GRU cell by improving the update gate and incorporating exponential linear unit activation. Additionally, SoftMax is replaced with Softsign activation in the output layer, which led to improved prediction accuracy of 95% with faster convergence compared to other sequential models. Another research [24] also achieved satisfactory recognition of about 99% through an RNN-based approach to address the issue of frame dependencies by using GRU, which outperformed LSTM and Bi-directional LSTM.

Table 1 shows the comparative analysis of related works.Existing research on sign language recognition, particularly on Pakistani Sign Language (PSL), reveals significant gaps in available

Table 1: Related work on sign language recognition

Related Work						
Sign Language	**Single / Double Handed**	**Static / Dynamic**	**Manual / Non-Manual**	**Alphabets / Numbers / Words**	**Recognition Model**	**Accuracy**
PSL [13]	Single	Static	Manual	Alphabets	SVM	89%
PSL [14]	Single	Static	Manual	Alphabets	SVM	80% - 90%
PSL [15]	Single	Static	Manual	Numbers	SVM, KNN, Random Forest	88%, 90%, 84%
PSL [16]	Single	Both	Manual	Alphabets	SVM	97.80%
PSL [17]	Both	Dynamic	Both	Words	C3D, I3D, TSM	93.33%
PSL [18]	Single	Both	Manual	Alphabet, Words	CNN	98.76%
CSL [19]	Both	Dynamic	Manual	Words	LSTM	90%
LSF [20]	Both	Dynamic	Manual	Words	LSTM	98.8%
GSL [21]	Both	Dynamic	Both	Words	3DCNN+ConvLSTM	98.5%
ISL [22]	Both	Dynamic	Manual	Words	SVM	99%
ISL [23]	Both	Dynamic	Both	Words	GRU	95%
ESL [24]	Both	Dynamic	Both	Words, Phrases	GRU, LSTM, BLSTM	99%

datasets suitable for training recognition systems. Although some datasets exist, they often lack sufficient variety and detail, especially at the word level. This study addresses this gap by introducing the first dataset containing seven PSL word-level pronouns. Different from previous sources, this dataset integrates manual and non-manual features without background restriction, providing a more comprehensive and representative training set for sign language recognition. Furthermore, the integration of the PKSMLNM dataset further improves the effectiveness of the training data, filling a critical gap in available resources for PSL identification.

In the domain of dynamic sign language recognition for Pakistan Sign Language (PSL), there is a notable lack of research focusing on processing sequential and temporal dependencies, particularly concerning double-handed gestures. Mutually, significant efforts have been made globally in various languages, using sophisticated model architectures for pose estimation and handcrafted features through deep learning techniques. However, within PSL research, pose estimation remains inadequately explored. Moreover, the accessibility of publicly available datasets poses a considerable challenge, as their generation is time-consuming, requiring sufficient computational resources and often involving multiple contributors, raising privacy concerns. Although several datasets exist for fingerspelling recognition in PSL, to our knowledge, only one dataset for word-level recognition was identified in local databases. To address these gaps, we propose methodologies to advance research efforts towards PSL recognition.

3. Proposed Methodology

We propose a method for real-time dynamic isolated sign language recognition of PSL as a contribution to word-level recognition within the research on PSL to bridge the gap of communication for the hearing impaired. We incorporate both manual and non-manual features with a comparison of RNN-based models; we evaluate our GRU and LSTM architectures' effectiveness in processing sequential data by leveraging multiple layers with additional dense layers for classification by utilizing several optimizers and augmentation techniques to assess the ability of models in processing two extensive datasets within comparison in a 3D space.

3.1. Data Collection

In our experiment, we have used PkSLMNM [11], which is a publicly available dataset, and our own dataset. This dataset comprises seven basic expressions, also termed as adjectives, such as bad, best, glad, scared, sad, surprised, and stiff, with 100 samples for each. Figure 1 shows the sample of 'Best' sign from our dataset. Due to the lack of word-level dynamic datasets, we extended our efforts towards isolated sign language recognition and introduced a dataset, contributing as the first pronouns-based dataset for PSL. We gained insights into the gestures through the PSL dictionary of an available PSL gestures learning resource, known as the Pakistan Sign Language application. Six basic pronouns were used, including he, she, me, you, this, and we, captured through a Sony A6100 camera. Our dataset was precisely created within our academic institution, i.e., University of Management and Technology in Sialkot, with consented contributions from 15 students with multiple static backgrounds with no objects around, such as people moving. Each student was first taught the sign and placed within 5 feet distance from the camera. We instructed the students throughout this process to perform one sign accurately with the right and left hand once, respectively. Signers may use their dominant hand for performing a gesture. To further address the generalization capabilities of the model, we incorporated both hands. Each video was recorded at 25 frames per second and comprised an average duration of 2 seconds. A total of 180 videos were generated, and both PkSLMNM and our datasets were selected for further processing. A snapshot of basic pronouns from our dataset is shown in Figure 2.

3.2. Preprocessing

The video data was further preprocessed by removing duplicate videos and cropping the frames.

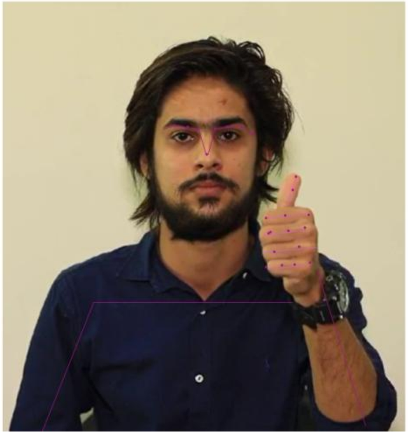


Figure 1: Mediapipe holistic landmarks and keypoints for pose, hands, and facial joints

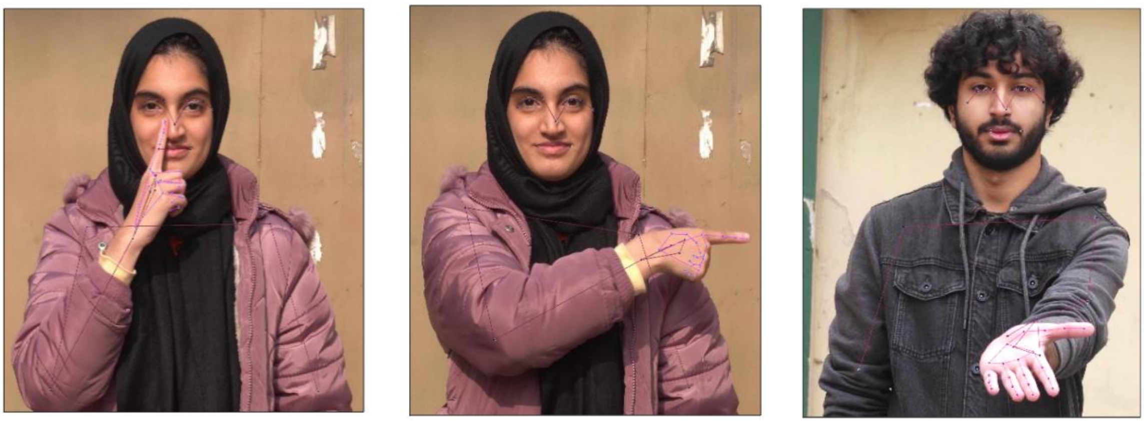


Figure 2: Sample of sign 'She' and 'This' in PSL of our dataset with holistic detection

Noise reduction was applied to compensate camera's poor stabilization to ensure each video consisted of consistent frames. We further preprocessed the PkSLMNM dataset due to its inconsistent frames, as it may pose additional challenges due to the intricacies of frame-by-frame processing for our classification model. The resolution was reduced by a factor of 2 in both dimensions to 960×540 to reduce the computational load. Further, data augmentation was applied to our training dataset using three techniques: scaling by ±0.1, rotation by ±10, and flipping horizontally by 1. Table 2 shows the data augmentation approach.

Table 2: Data augmentation applied to 180 original videos: scaling (±0.1) → 360 videos, rotation (±10°) → 360 videos, horizontal flipping → 180 videos.

Augmentation		
Techniques	**Videos**	
	Original x Augmented	**Total Videos**
Scaling	180 x (+0.1, -0.1)	360
Rotating	180 x (+10, -10)	360
Flipping	180 (horizontal by 1)	180

3.3. Feature Extraction

We used holistic detection to draw landmarks on our participants and PkSLMNM videos, which was done for key elements like pose, left hand, right hand, and facial landmarks within each frame of the video. Each landmark contains three-dimensional coordinates (X, Y, Z) indicating its position in the image, along with a visibility score indicating the confidence level of the detection. These landmarks serve as representations of distinct structural locations within the detected hands, pose, and face. It further eases the complexity by involving a hand tracking feature, as it has 21 hand landmarks that depict various points on the hand, encompassing fingertips, joints, and the palm. Pose estimation involves detecting 33 landmarks for essential body positions like shoulders, elbows, eyes, mouth, and 468 facial landmarks. MediaPipe landmark enables accurate localization and tracking of these pivotal points, facilitating our application, as shown in Figure 3. Further, these keypoints were extracted from the pose, left and right hands, and face. Three coordinates, x, y, and z, and visibility were considered within pose, hence 33×4 equals 132 keypoints. For hands and face, only three coordinates, x, y, and z, were considered; hence, 21*3 for the left hand and 21×3 for the right hand equals 126 keypoints and 468×3 equals 1404. The visibility parameter for pose keypoints, but not for hands or face, was selected due to domain-specific considerations in sign language recognition. In typical sign language videos, which are recorded in controlled environments like our dataset, the hands and face are invariably in the foreground, leading to high detection confidence and minimal variance in visibility scores (often close to 1.0). So, including visibility for these modules would bring redundant features rather than informational gain, which further increases computational complexity and risk of overfitting for our model. Whereas for pose keypoints, visibility score facilitates in detecting accurate spatial association and error handling in the feature set, as it involves consistent body positioning, clothing, or slight movements. So, the total number of features extracted had a fixed duration of 55 frames, comprising 1662 features. These landmarks are flattened into an array serving as input for our model. In cases where a landmark is not detected, the array is padded with zeroes.

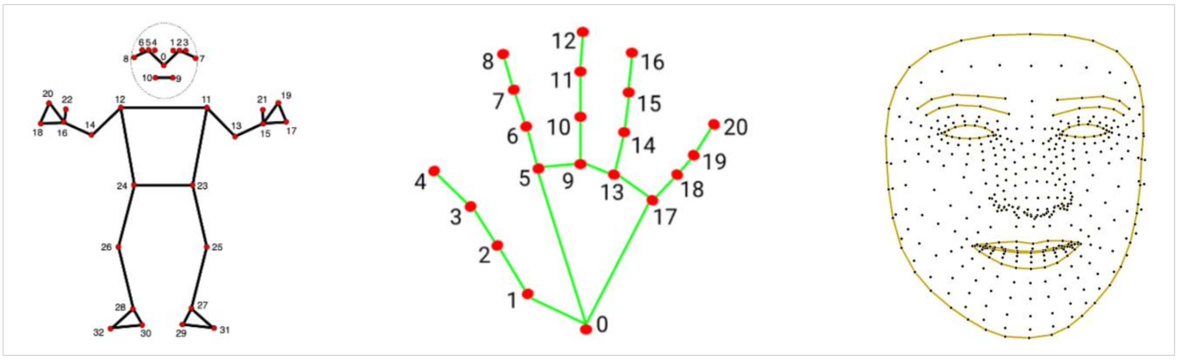


Figure 3: Sample of 'Best' sign in PSL from PkSLMNM dataset

Then, within each sequence, a frame-by-frame analysis is conducted, and the relevant keypoint information is extracted from the stored data, reflecting the spatial coordinates of key features. A label map is crafted, associating each label with specific numerical values, expediating the conversion of gesture labels into their numerical representations. For each sequence, the corresponding numerical label is appended to the labels list, aligning the temporal data with its categorical representation, and further undergoes transformation into numpy arrays, and sequences array encapsulates organized keypoint data.

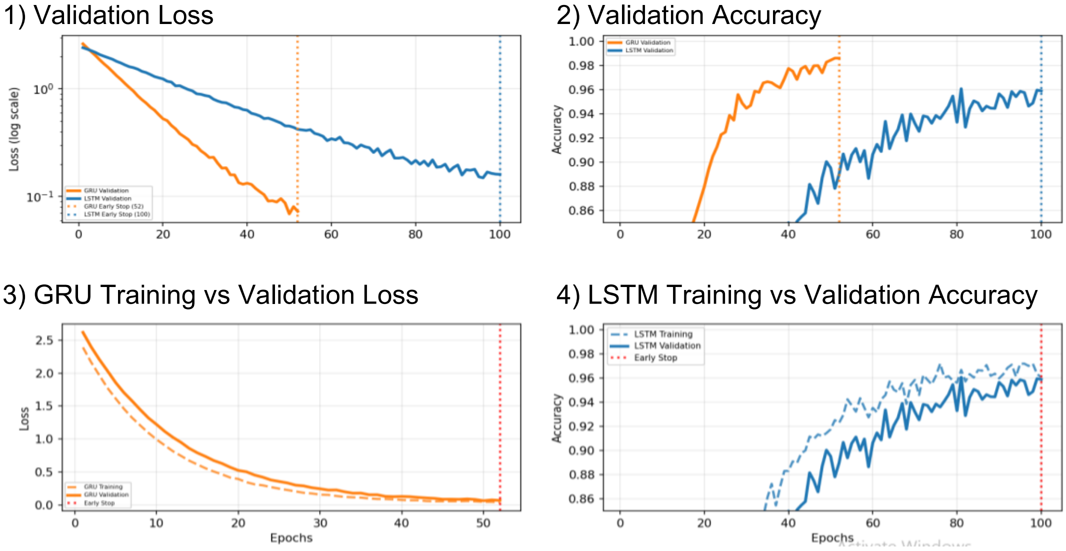
3.4. Classification

The sequential model is constructed using the TensorFlow Keras API, featuring 3 LSTM layers with 256, 128, and 64 units, all with ReLU activation return sequences as true for first two layers, followed by a dropout layer, & dense layer with 128 units and ReLU activation, and finally an output dense layer with softmax activation corresponding to the number of pronoun classes. The input shape is explicitly set as (56, 1662), encapsulating both sequence length and features extracted from each frame with a ReLU activation function to introduce non-linearity. Data partitioning entails the division of loaded sequences and labels into training and testing sets. An 80/20 stratified train/validation split was performed using scikit-learn's train_test_split (stratify=y, random_state=42) to preserve class distribution and for full reproducibility. The model undergoes compilation using the Adam and Stochastic Gradient Descent (SGD) optimizer and categorical cross-entropy loss, with dropout rates of 0.4 and 0.3 at default, and multiple learning rates for testing with a batch size of 32 and 64 as given in Table 3. The resulting model is saved in an h5 file. The same architecture was applied for GRU. Both models were trained with callbacks for early stopping with a maximum of 200 and 300 epochs. Evaluation unfolds on the training set, involving predictive modeling, the computation of multilabel confusion matrices, and the derivation of accuracy scores. Additional provisions are made for loading weights and conducting supplementary predictions on the test set. Both models' results are compared, and accuracies are given in the following Table 3.

Table 3: Experimental hyperparameter configurations used for our systematic model evaluation

No#	Hyperparameters				
	Dropout	Learning Rate	Optimizer	Validation Split	Early Stopping
Experiment 1	0.4 & 0.3	0.001	Adam & SGD	80% train+val, 20% test	200 (patience=10, monitoring val_loss)
Experiment 2	0.3 & 0.2	0.8			300 (patience=8, monitoring val_loss)

The training dynamics comparison as shown in Figure 4 indicates that the GRU's 48% faster convergence and superior efficiency (52 vs 100 epochs) with maintained generalization across validation metrics. The bottom panels denote stable learning without overfitting for both architectures.


1) Validation Loss

2) Validation Accuracy

3) GRU Training vs Validation Loss

4) LSTM Training vs Validation Accuracy

Figure 4: Training dynamics comparison

4. Results and Discussion

From our results on our pronouns dataset, we concluded that GRU performed well with Adam optimizer, whereas LSTM also delivered satisfactory results. The former optimizers were used to evaluate optimization efficacy under adaptive and fixed learning-rate regimes. Adam, with its parameter-specific adaptive learning rates, proved effective for navigating the complex loss landscape of temporal models, leading to faster and more reliable convergence. Whereas SGD's performance was highly sensitive to its fixed learning rate, which resulted in slower convergence and a greater tendency to settle in sub-optimal minima that indicating its lower accuracy across all model and dataset configurations. Through our training process with both datasets, GRU was computationally efficient with a dropout of 0.4 and a default learning rate of 0.001, and was trained faster than LSTM. GRU early-stopped at 52 epochs while LSTM stopped at 100 epochs with (patience=10, monitoring val_loss). Categorical cross-entropy loss, tailored for multi-class classification problems, is chosen to measure dissimilarity between true labels and predicted probabilities, where all correct predictions were made on our dataset using GRU with Adam optimizer. Categorical accuracy is used as a metric to calculate the accuracy of the model's predictions by comparing predicted class labels to true class labels, and GRU achieved an impressive and higher final validation and test accuracies of 98.61% to LSTM's 96.13%. In Table 4 and Figure 5, the results show that the GRU model outperforms the LSTM model on all datasets and optimization algorithms.

Table 4: Comparison of GRU and LSTM Accuracy

	GRU		**LSTM**	
	Adam	**SGD**	**Adam**	**SGD**
Pronouns Dataset	94.71%	91.34%	92.44%	89.42%
PkSLMNM Dataset	93.25%	89.52%	88.17%	90.32%
Pronoun Augmented Dataset	98.61%	94.1%	96.13%	93.5%

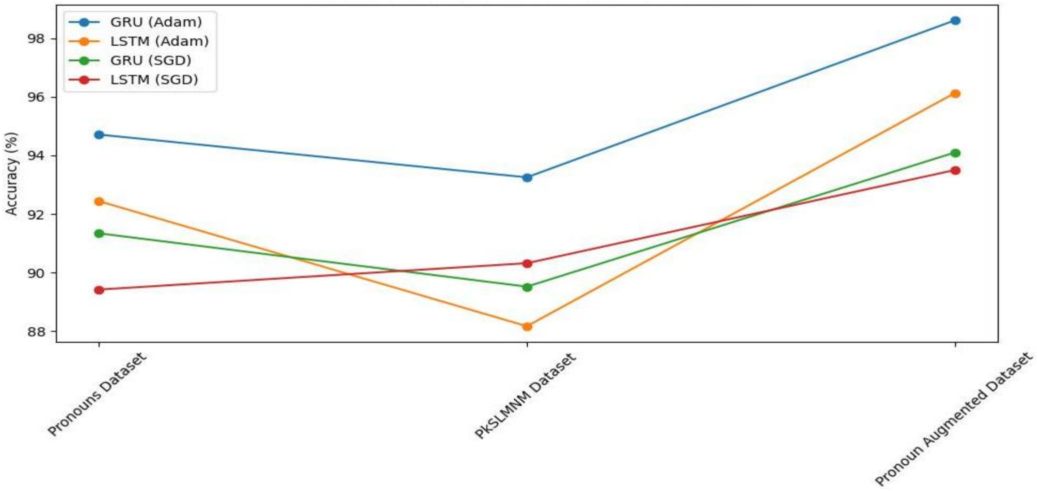


Figure 5: Accuracy Comparison of the GRU and LSTM models of Pronouns (Original & Augmented), and PkSLMNM datasets

Figure 5 depicts the consistent out-performance of GRU over LSTM, which can be attributed to its basic architectural advantages for this task. GRU's blueprint highlights a simplified gating mechanism (update and reset) in contrast with LSTM's three gates (input, forget & output). The limited parameters in GRU make it less inclined towards overfitting, especially on datasets of moderate size, enabling faster, & more efficient training; a finding directly supported by GRU's earlier convergence (52 vs. 100 epochs). Moreover, GRU's update gate covers effectively for LSTM's input and forget gates, which allows it to capture long-range temporal dependencies in sign language gestures without unnecessary complexity. Whereas pose keypoints demand sequential data, where information flow is more streamlined than in raw video, so, this streamlined architecture is sufficiently powerful to model the essential dynamics. The result is a model that generalizes better from our training data to unseen signers and environmental variations, as confirmed by its higher test accuracy.

5. Conclusions

In the domain of dynamic sign language recognition, a notable gap exists in processing sequential and temporal dependencies, particularly concerning Pakistan Sign Language (PSL). While extensive efforts have been made globally, including sophisticated model architectures, PSL research remains largely unexplored in pose estimation, exacerbated by the scarcity of publicly available datasets. To address these gaps, we propose a novel sign language pose estimation-based recognition system tailored for PSL. Our approach introduces the first dataset comprising seven word-level pronouns, including both manual and non-manual features, without background constraints. Additionally, integration of the PKSMLNM dataset improves the potency of our training data. Using Mediapipe Holistic, a comprehensive feature extraction is done, while LSTM and GRU models effectively capture the temporal dependencies within these extracted features. Our study illuminates effective strategies for handling large datasets in dynamic gesture analysis and tackles computational resource challenges. Evaluation results indicate GRU's superiority over LSTM, demonstrating computational efficiency and accelerated training. Specifically, the Adam optimizer proves effective for GRU, yielding impressive accuracy. Ultimately, GRU emerges as a promising model for efficient dynamic sign language recognition, despite limitations in MediaPipe detection distance. In the future, we can improve the accuracy of the model by integrating multimodal features, such as sign language gestures and expressions, by using deep learning advanced algorithms.

6. References

[1] "International Day of Sign Languages,"—United Nations.

[2] Pakistan Association of the Deaf, "Pakistan Association of the Deaf Deaf Statistics," Pakistan Association of the Deaf.

[3] Elakkiya R., "Machine learning based sign language recognition: a review and its research frontier," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 7, pp. 7205-24, July 2021.

[4] Zeyu, Huailing Li, and Jianping Chai Liang, "Sign language translation: A survey of approaches and techniques," *Electronics*, vol. 12, no. 12, p. 2678, June 2023.

[5] Chatzis T, Papastratis I, Stergioulas A, Papadopoulos GT, Zacharopoulou V, Xydopoulos GJ, Atzakas K, Papazachariou D, Daras P Adaloglou N, "A comprehensive study on sign language recognition methods," arXiv preprint arXiv:2007, vol. 2, no. 2, p. 12530, 2020.

[6] Q., Sun, L., Han, C., Guo, J Gao, "American sign language fingerspelling recognition using RGB-D and DFANet," *Proc. China Automation Congress (CAC)*, pp. 3151-3156, November 2022.

[7] S.M., Chen, Y., Li, S., Shi, X., Zheng, J Kamal, "Technical Approaches to Chinese sign language processing: a review," *IEEE Access,* 7, pp. 96926–96935, July 2019.

[8] O. Koller, S. Hadfield, and R. Bowden N. C. Camgoz, "Sign language transformers: Joint end-to-end sign language recognition and translation," *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 10023–10033, June 2020.

[9] Ulrich Von Agris and Karl-Friedrich Kraiss, "Towards a video corpus for signer-independent continuous sign language recognition Gesture in Human-Computer Interaction and Simulation," *International Gesture Workshop*, 2007.

[10] Hao, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li Zhou, "Improving sign language translation with monolingual data by sign back-translation," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1316-1325, 2021.

[11] Rizvi S Javaid S, "A Novel Action Transformer Network for Hybrid Multimodal Sign Language Recognition," *Computers, Materials & Continua*, vol. 75, no. 1, April 2023.

[12] R., Kiani, K. and Escalera, S Rastgoo, "Sign language recognition: A deep survey," *Expert Systems with Applications*, no. 164, p.113794, 2021.

[13] M. S. Shah, W. Akram, A. Manzoor, R. O. Mahmoud and D. S. Abdelminaam F. Shah, "Sign Language Recognition Using Multiple Kernel Learning: A Case Study of Pakistan Sign Language," *IEEE Access*, vol. 9, no. 4, pp. 67548-67558, May 2021.

[14] Razzaq, I. Ahmad Baig, A. Hussain, S. Shahid, and T.-ur Rehman A. Imran, "Dataset of Pakistan Sign Language and Automatic Recognition of Hand Configuration of Urdu Alphabet through Machine Learning," *Data in Brief*, vol. 36, p. 107021, June 2021.

[15] H., Hira, M. R., Syed, S. A., Ullah, R., Asif, M., Khan, M., Mujeeb, A. A., & Khan, A. H. Zahid, "A computer vision-based system for recognition and classification of Urdu sign language dataset.," *PeerJ Computer Science*, vol. 8, p. e1174, Dec 2022.

[16] M. S., Munaf, S. M., Azim, F., Ali, S., & Khan, S. J. Mirza, "Vision-based Pakistani sign language recognition using bag-of-words and support vector machines.," *Scientific Reports*, vol. 12, p. 21325, June 2022.

[17] Wali Hamza HM., "Pakistan sign language recognition: leveraging deep learning models with limited dataset.," *Machine Vision and Applications*, vol. 34, no. 5, p. 71, Sep 2023.

[18] S. O. Rana, N. M. Ansari, R. Iqbal, T. Tariq, M. U. Awan, and A. Waqar I. M. Khan, "Design and implementation of CNN for sign language recognition," *Journal of Jilin University (Engineering and Technology Edition)*, vol. 41, p. 135-145. Oct 2022.

[19] Zhou W, Li H. Liu T, "Sign language recognition with long short-term memory," *In 2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2871-2875, Sep 2016.

[20] A., Lopez-Cabrera, V., Rangel Teran-Quezada, "Sign-to-Text Translation from Panamanian Sign Language to Spanish in Continuous Capture Mode with Deep Neural Networks," **Big Data and Cognitive Computing**, vol. 8, no. 3, p. 25, Feb 2024.

[21] D. R. Fathy E. K. Elsayed, "Semantic Deep Learning to Translate Dynamic Sign Language," *International Journal of Intelligent Engineering & Systems*, vol. 14, no. 1, Jan 2021.

[22] Tayade Halder A, "Real-time Vernacular Sign Language Recognition using Mediapipe and Machine Learning," *International Journal of Research Publication and Reviews*, vol. 2582, vol. 2, no. 5, p. 9-17, 2021.

[23] Olimov, Naik SM, S. Kim, Park KH, J. Kim B. Subramanian, "An integrated mediapipe-optimized GRU model for Indian sign language recognition," *Scientific Reports*, vol. 12, no. 1, p. 11964, Jul 2022.

[24] R. Wadie, A. K. Attia, A. M. Asaad, A. E. Kamel, S. O. Slim, M. S. Abdallah MS, Y. I. Chou G. H. Samaan, "Mediapipe's landmarks with RNN for dynamic sign language recognition.," *Electronics*, vol. 11, no. 19, p. 3228, Oct 2022.