Jurnal Informatika dan Rekayasa Perangkat Lunak ISSN 2656-2855
Vol. 6, No. 2, September 2024, Hal. 427-432 e-ISSN 2685-5518

Pengenalan Gestur Bahasa Isyarat Indonesia dengan Mediapipe Keypoints

Febrian Murti Dewanto[1*], **Aris Tri Jaka Harjanta**[2], **Noora Qotrun Nada**[3], **Bambang Agus Herlambang**[4]
[1,2,3,4] Program Studi Informatika, Fakultas Teknik dan Informatika, Universitas PGRI Semarang
[*]Email: febrianmd@upgris.ac.id

Abstrak

Kesulitan dalam berkomunikasi merupakan hambatan bagi teman tuli yang tidak bisa mempelajari bahasa secara lisan atau memperoleh kemampuan bicara yang biasa. Pengembangan teknologi pengenalan gestur bahasa isyarat merupakan langkah penting untuk meningkatkan aksesibilitas dan integrasi sosial bagi komunitas tuli. Penggunaan MediaPipe Holistic Keypoints dan teknik deep learning memberikan potensi yang signifikan dalam mengenali dan memahami gestur bahasa isyarat. Tujuan utama penelitian ini adalah mengklasifikasikan gestur Bahasa Isyarat Indonesia (BISINDO) menggunakan MediaPipe Holistic Keypoints dan pendekatan deep learning untuk mengidentifikasi kata dasar dalam bahasa isyarat. Dengan ekstraksi fitur menggunakan mediapipe holistic dan mengirimnya ke model LSTM 6 hidden layer dengan 70:30 split train test dan 250 epoch dihasilkan akurasi 68 %. Hal ini dikarenakan terbatasnya jumlah dataset yang diambil untuk penelitian.

***Kata kunci**: Bahasa Isyarat, LSTM, MediaPipe, Rekognisi*

Abstract

Difficulty in communication is an obstacle for deaf friends who cannot learn the language orally or acquire normal speech skills. The development of sign language gesture recognition technology is an important step to improve accessibility and social integration for the deaf community. The use of MediaPipe Holistic Keypoints and deep learning techniques provides significant potential in recognizing and understanding sign language gestures. The main objective of this study is to classify Indonesian Sign Language (BISINDO) gestures using MediaPipe Holistic Keypoints and a deep learning approach to identify basic words in sign language. By extracting features using mediapipe holistic and sending them to the LSTM 6 hidden layer model with 70:30 split train test and 250 epochs, an accuracy of 68% was produced. This is due to the limited number of datasets taken for the study.

***Keywords**: Sign Language, LSTM, MediaPipe, Recognition*

PENDAHULUAN

Kesulitan dalam berkomunikasi menjadi kendala bagi teman tuli yang tidak dapat mempelajari bahasa dalam bentuk verbal atau memperoleh kemampuan berbicara secara normal. Salah satu cara mereka untuk berkomunikasi adalah menggunakan bahasa isyarat. Komunikasi bahasa isyarat adalah sarana utama untuk berinteraksi dengan dunia. Pengenalan Bahasa Isyarat atau *Sign Language Recognition* (SLR) bertujuan untuk mengatasi kesenjangan antara pengguna bahasa isyarat dan orang lain dengan mengenali isyarat-isyarat dari video yang diberikan. Ini adalah penelitian yang penting namun menantang, karena bahasa isyarat dilakukan dengan gerakan tangan yang cepat dan kompleks, postur tubuh, dan bahkan ekspresi wajah(Jiang et al., 2021). Terdapat berbagai tantangan dalam tugas pengenalan bahasa isyarat, pada tahap ini belum mungkin untuk merancang alat (*SLR tools*) yang mencapai akurasi 100% untuk kosakata besar atau banyak dalam bahasa isyarat. Potensi penerapan alat pengenalan bahasa isyarat yang efektif. Misalnya, dapat menerjemahkan siaran yang menyertakan bahasa isyarat, membuat perangkat yang bereaksi terhadap perintah bahasa isyarat, atau bahkan merancang sistem untuk membantu penyandang disabilitas dalam melakukan pekerjaan rutin. Secara khusus, *Deep Neural Network* (DNN) telah muncul sebagai aset yang berpotensi menjadi terobosan bagi para peneliti, dan dampak penuh penerapannya

terhadap masalah SLR kedepannya(Al-qurishi et al., 2021).

Di Indonesia terdapat dua bahasa isyarat yaitu Sistem Bahasa Isyarat Indonesia (SIBI) dan Bahasa Isyarat Indonesia (BISINDO)(Moetia Putri and Fuadi, 2020). Ada banyak perbedaan antara SIBI dan BISINDO, salah satunya diadopsi dari ASL (*American Sign Language*), yang ini disebut SIBI. Namun SIBI dan BISINDO masih digunakan di Indonesia. Meskipun demikian, SIBI telah disetujui oleh pemerintah Indonesia, dan SIBI digunakan di sekolah dan untuk belajar, namun sebagian besar penyandang tunarungu di Indonesia lebih banyak menggunakan BISINDO dalam aktivitas kehidupannya (Aljabar and Suharjito, 2020).

Dengan hadirnya Framework MediaPipe yang berfungsi sebagai ekstraksi fitur yang relevan dan klasifikasi data untuk menentukan *keypoints* paling mungkin sedang dipresentasikan. Semakin banyak penelitian yang menggunakan Framework MediaPipe untuk ekstraksi fitur dalam SLR seperti (Akshit Tayade and Halder, 2021; Badarinath and Shamitha, 2023; Bora et al., 2023; Shamitha S H and Badarinath K, 2023). Misalnya pada penelitian sistem pengenalan Bahasa Isyarat India (ISL) *real-time* untuk 24 sinyal dinamis menggunakan kerangka Mediapipe dan jaringan LSTM. Metode yang diusulkan dalam penelitian melibatkan pelatihan LSTM untuk membedakan berbagai gesture menggunakan kumpulan data yang dibuat dari 24 tanda isyarat dinamis. Untuk menyelesaikan pembuatan kumpulan data, model Holistik kerangka Mediapipe yang telah dilatih sebelumnya digunakan sebagai ekstraktor fitur(Badarinath and Shamitha, 2023).

Dari berbagai penelitian pengenalan bahasa isyarat yang memanfaatkan Mediapipe tersebut, kebutuhan untuk mengembangkan model yang bisa mengklasifikasi gestur bahasa isyarat untuk memfasilitasi komunikasi antara masyarakat dan teman Tuli di Indonesia akan dilakukan pada penelitian ini.Perbedaan penelitian ini dengan sebelumnya adalah menggunakan dataset private dengan arsitektur yang berbeda antara lain menggunakan semua keypoints dari Mediapipe dan model RNN LSTM.

KAJIAN TEORITIS

Rekognisi Bahasa Isyarat

Pengenalan atau rekognisi bahasa isyarat melibatkan analisis dan interpretasi gerakan dan tanda-tanda tangan untuk menerjemahkan bahasa isyarat ke dalam teks atau aksi lainnya. Tantangan yang mungkin dihadapi termasuk variasi dalam gerakan isyarat, kecepatan, dan faktor lingkungan lainnya yang dapat mempengaruhi akurasi sistem pengenalan bahasa isyarat. Dibandingkan dengan pengenalan tindakan konvensional, Pengenalan Bahasa Isyarat merupakan masalah yang lebih menantang. Pertama, bahasa isyarat memerlukan gerakan tubuh keseluruhan dan gerakan tangan/ lengan yang halus untuk menyatakan maknanya secara jelas dan akurat. Ekspresi wajah juga dapat digunakan untuk menyatakan emosi.

Gerakan serupa bahkan dapat memiliki berbagai makna tergantung pada jumlah pengulangan. Kedua, penutur bahasa isyarat yang berbeda dapat melakukan bahasa isyarat dengan cara yang berbeda (kecepatan, lokalisme, tangan kiri, tangan kanan, bentuk tubuh), menjadikan SLR lebih menantang. Mengumpulkan lebih banyak sampel dari sebanyak mungkin penutur bahasa isyarat diinginkan namun mahal. Metode SLR tradisional utamanya menggunakan fitur yang dibuat secara manual seperti HOG (Qiang Zhu et al., 2006) dan SIFT yang dikombinasikan dengan klasifikasi konvensional seperti kNN sudah dilakukan, karena bahasa isyarat dilakukan dengan gerakan tangan yang cepat dan kompleks, postur tubuh, dan bahkan ekspresi wajah. Baru-baru ini, estimasi pose / *pose estimator* berbasis kerangka tubuh / *skeleton aware* semakin menarik perhatian karena kemandiriannya terhadap variasi subjek dan latar belakang.(Jiang et al., 2021)

MediaPipe

Para peneliti Google dengan BlazePose, sebuah arsitektur jaringan saraf konvolusional yang ringan untuk estimasi pose manusia yang dirancang khusus untuk inferensi waktu nyata pada perangkat seluler. Selama inferensi, jaringan menghasilkan 33 titik kunci tubuh untuk satu orang dan berjalan dengan kecepatan lebih dari 30 frame per detik pada ponsel Pixel 2. Hal ini membuatnya sangat cocok untuk penggunaan waktu nyata seperti pelacakan kebugaran dan pengenalan bahasa isyarat.(Bazarevsky and Zhang, 2020)

Dibandingkan dengan sebagian besar solusi estimasi pose yang ada yang mendeteksi titik kunci menggunakan heatmap, solusi berbasis pelacakan kami memerlukan penyelarasan pose awal. Dataset Blazepose pada kasus-kasus di mana seluruh tubuh orang terlihat, atau di mana titik kunci pinggul dan bahu dapat diannotasi dengan keyakinan.

Untuk memastikan bahwa model mendukung oklusi berat yang tidak ada dalam dataset, menggunakan augmentasi yang mensimulasikan oklusi yang substansial. Dataset pelatihan terdiri dari 60 ribu gambar dengan satu atau beberapa orang dalam adegan dengan pose umum dan 25 ribu gambar dengan satu orang dalam adegan melakukan latihan kebugaran. Semua gambar ini diannotasi oleh manusia. Arsitektur jaringan saraf kami pada komponen estimasi pose sistem kami memprediksi lokasi semua 33 titik kunci orang, dan menggunakan proposal penyelarasan orang yang disediakan oleh tahap pertama dari pipeline.(Bazarevsky and Zhang, 2020) Framework yang terus dikembangkan peneliti google adalah MediaPipe dimana menggunakan arsitektur jaringan saraf konvolusional, yang bisa melakukan deteksi wajah, tangan dan pose estimasi dengan titik kunci /*keypoints* (Lugaresi et al., 2019).

Long Short Term Memory (LSTM)

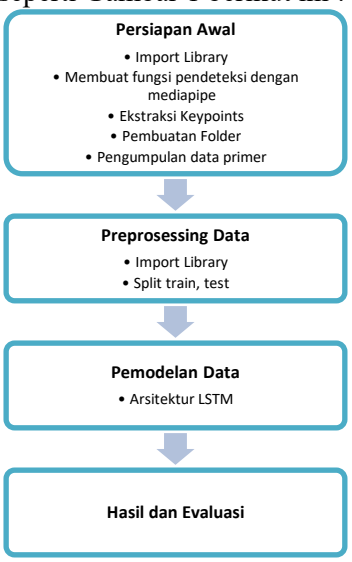
Long short-term memory (LSTM) merupakan pengembangan dari arsitektur *Recurrent Neural Network* (RNN). LSTM memiliki keunggulan dalam mengingat dan menyimpan informasi masa lampau serta mampu mempelajari suatu data yang bersifat sekuensial(Moetia Putri and Fuadi, 2020). Terdapat struktur dasar pada LSTM yaitu input layer, hidden layer, dan output layer. Terdapat dua fungsi aktivasi yang digunakan pada LSTM yaitu sigmoid dan tanh. Pada LSTM juga terdapat *memory cell* dan gerbang. Gerbang tersebut tersusun dari tiga gerbang yaitu *forget gate, input gate*, dan *output gate*. Persamaan metode LSTM dapat dilihat dibawah ini :

ft = σ (wf . [ht−1, xt] + bf) (1)
it = σ (wi . [ht −1, xt] + bi) (2)
Ct ~ = tanh (wc . [ht−1, xt] + bc) (3)
Ct = ft ∗ Ct−1 + it ∗ Ct ~ (4)
ot = σ (wt . [ht−1 , xt] + bo) (5)
ht = ot ∗ tanh (Ct) (6)

Dimana matriks w merupakan bobot, b merupakan nilai bias, dan ft , it , Ct ~, ot , ht merupakan keluaran dari forget gate, input gate, cell state, output gate, dan nilai output pada waktu (t).

METODE PENELITIAN

Penelitian dalam klasifikasi gestur bahasa isyarat dengan Mediapipe keypoints ini adalah seperti Gambar 1 berikut ini :



Gambar 1 Metode Penelitian

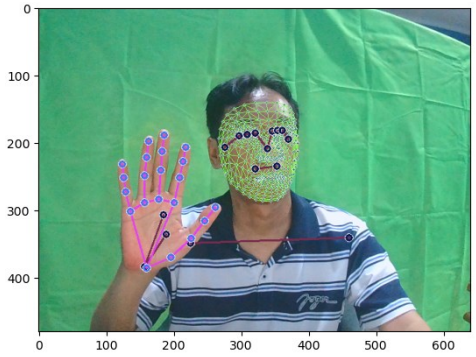
HASIL DAN PEMBAHASAN

Persiapan Awal

Pada tahap Persiapan Awal, yaitu:

1. Import Library kebutuhan dalam penelitian.
   - import cv2 (OpenCV untuk tugas Visi Komputer)
   - import numpy as np (array)
   - import os (Folder)
   - import time
   - import mediapipe as mp (Ekstraksi fitur keypoints)
2. Membuat fungsi pendeteksi landmarks tangan, wajah, dan badan.

Kamera webcam diuji coba akan merekam tangan (hand), wajah (face), dan badan(pose) menggunakan open cv dan mediapipe holistic. Landmark yang dihasilkan dalam satu frame gambar 21 keypoints pada setiap tangan, 468 keypoints pada wajah, dan 33 keypoints pada badan.



Gambar 2 Keypoints dengan Mediapipe

3. Proses ekstraksi keypoints

Keypoints yang telah diperoleh sebelumnya akan diekstraksi dengan menggabungkan nilai keypoints tersebut ke dalam array numpy.

Face = 468 x 3 (x,y, z) , Pose = 33 x4 (x,y,z,vis) , Hand = 21x 2 x 3 (x,y,z)

Shape perr frame setelah digabung (1662)

4. Persiapan pembuatan folder

Folder ini berguna sebagai tempat penyimpanan data kosakata isyarat yang telah direkam oleh webcam dan telah dideteksi menggunakan keypoints mediapipe holistic.

Folder kelas diberi nama actions disimpan dalam np.array(['aku', 'kamu', 'saya', 'dia', 'kita', 'kalian', 'apa', 'siapa', 'kenapa', 'berapa'])

5. Mengumpulkan keypoints

Proses perekaman gestur kosakata BISINDO dilakukan dengan menggunakan webcam dengan *dependent signer* merekam keypoints dari setiap gerakan kosakata pada BISINDO. Setiap kosakata memiliki 10 video sequence dan setiap sequences akan merekam sebanyak 30 frame. Keypoints yang telah direkam akan tersimpan otomatis kedalam folder yang sudah dibuat sebelumnya. Dengan masing-masing video 10 file npy.

Preprosesing Data

1. Import Library kebutuhan preprosesing data.
   from sklearn.model_selection import train_test_split
   from tensorflow.keras.utils import to_categorical
2. Memberi label data one hot encoding :{'aku': 0, 'kamu': 1, 'saya': 2, 'dia': 3, 'kita': 4, 'kalian': 5, 'apa': 6, 'siapa': 7, 'kenapa': 8, 'berapa': 9} Shape : Setelah digabung dengan label (100, 30, 1662)
3. Split train dan test

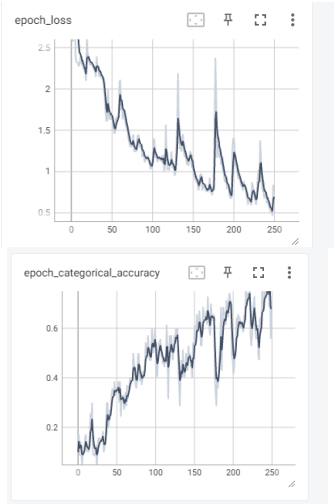
Ukuran 70% dan 30%, dengan stratify=y untuk mengimbangkan jumlah kelas test

Pemodelan data

1. Import library
   from tensorflow.keras.models import Sequential
   from tensorflow.keras.layers import LSTM, Dense
   from tensorflow.keras.callbacks import TensorBoard
2. Membuat model dengan Arsitektur sebagai berikut
   model = Sequential()
   model.add(LSTM(64, return_sequences=True, activation='relu', input_shape=(30,1662)))
   model.add(LSTM(128, return_sequences=True, activation='relu'))
   model.add(LSTM(64, return_sequences=False, activation='relu'))
   model.add(Dense(64, activation='relu'))
   model.add(Dense(32, activation='relu'))
   model.add(Dense(actions.shape[0], activation='softmax'))
3. Menentukan optimizer Adam
   model.compile(optimizer='Adam', loss='categorical_crossentropy', metrics=['categorical_accuracy'])
4. Melakukan Pemodelan dengan ujicoba 250 epoch
   model.fit(X_train, y_train, epochs=250, callbacks=[tb_callback])

Hasil dan Evaluasi

1. Hasil dari training model, seperti gambar yang dihasilkan pada tensorboard sebagai berikut:



Gambar 3 Loss dan Accuracy

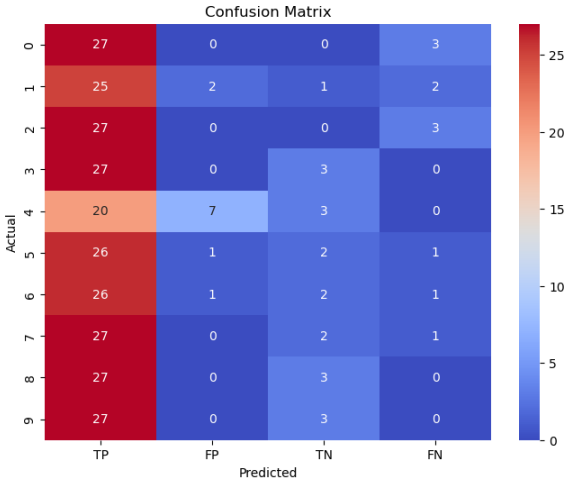
Dengan Summary Model parameter : "sequential_4"

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
lstm_12 (LSTM)              (None, 30, 64)            442112
lstm_13 (LSTM)              (None, 30, 128)           98816
lstm_14 (LSTM)              (None, 64)                49408
dense_12 (Dense)            (None, 64)                4160
dense_13 (Dense)            (None, 32)                2080
dense_14 (Dense)            (None, 10)                330
=================================================================
Total params: 596906 (2.28 MB)
Trainable params: 596906 (2.28 MB)
Non-trainable params: 0 (0.00 Byte)
```

2. Hasil categorical accuracy 68,57%
   Epoch 250/250
   3/3
   [==============================]
   - 1s 232ms/step - loss: 0.7667 - categorical_accuracy: **0.6857**
3. Evaluasi dengan Confusion Matriks multilabel.
   Untuk menghitung akurasi dari confusion matrix menggunakan rumus berikut:
   Accuracy=(TP+TN)/(TP+TN+FP+FN)

Di mana:

- TP = True Positives (jumlah prediksi yang benar positif)
- TN = True Negatives (jumlah prediksi yang benar negatif)
- FP = False Positives (jumlah prediksi yang salah positif)
- FN = False Negatives (jumlah prediksi yang salah negatif)

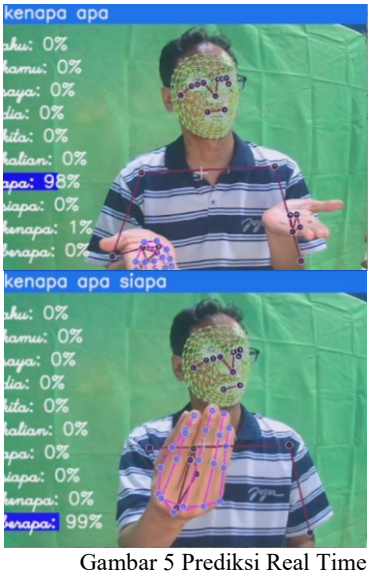


Gambar 4 Confusion Matrix

Akurasi mengukur seberapa sering model mengklasifikasikan dengan benar data yang diamati. Semakin tinggi nilai akurasi, semakin baik kinerja model.

1. Akurasi untuk setiap data:
2. Data ke-1: 0.9
3. Data ke-2: 0.87
4. Data ke-3: 0.9
5. Data ke-4: 1.0
6. Data ke-5: 0.77
7. Data ke-6: 0.93
8. Data ke-7: 0.93
9. Data ke-8: 0.97
10. Data ke-9: 1.0
11. Data ke-10: 1.0

4. Prediksi secara realtime
   Selanjutnya model tersebut diprediksi secara realtime, dengan hasil sebagai berikut :



Gambar 5 Prediksi Real Time

SIMPULAN

Penggunaan *MediaPipe Holistic Keypoints* dan teknik deep learning memberikan potensi yang signifikan dalam mengenali dan memahami gestur bahasa isyarat. Tujuan utama penelitian ini adalah mengklasifikasikan gestur Bahasa Isyarat Indonesia (BISINDO) menggunakan *MediaPipe Holistic Keypoints* dan pendekatan deep learning untuk mengidentifikasi kata dasar dalam bahasa isyarat. Dengan ekstraksi fitur menggunakan mediapipe holistic dan mengirimnya ke model LSTM 6 hidden layer dengan 70:30 *split train test* dan 250 epoch dihasilkan akurasi 68 %. Saran kedepannya adalah menambah jumlah dataset yang diambil untuk penelitian. Kemudian untuk penelitian selanjutnya adalah ujicoba menggunakan model selain LSTM seperti GRU dan menambah jumlah dataset.

DAFTAR PUSTAKA

Akshit Tayade, Halder, A., 2021. Real-time Vernacular Sign Language Recognition using MediaPipe and Machine Learning. https://doi.org/10.13140/RG.2.2.32364.03203

Aljabar, A., Suharjito, S., 2020. BISINDO (Bahasa Isyarat Indonesia) Sign Language Recognition Using CNN and LSTM. Adv. Sci. Technol. Eng. Syst. J. 5, 282–287. https://doi.org/10.25046/aj050535

Al-qurishi, M., Khalid, T., Souissi, R., 2021. Deep Learning for Sign Language Recognition : Current Techniques , Benchmarks , and Open Issues. IEEE Access PP, 1. https://doi.org/10.1109/ACCESS.2021.3110912

Badarinath, Shamitha, 2023. Sign Language Recognition utilizing LSTM and Mediapipe for Dynamic Gestures of ISL. Int. J. Multidiscip. Res. 5, 6868. https://doi.org/10.36948/ijfmr.2023.v05i05.6868

Bazarevsky, V., Zhang, F., 2020. BlazePose : On-device Real-time Body Pose tracking.

Bora, J., Dehingia, S., Boruah, A., Chetia, A.A., Gogoi, D., 2023. Real-time Assamese Sign Language Recognition using MediaPipe and Deep Learning. Procedia Comput. Sci. 218, 1384–1393. https://doi.org/10.1016/j.procs.2023.01.117

Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., Fu, Y., 2021. Skeleton Aware Multi-modal Sign Language Recognition.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M.G., Lee, J., Chang, W.-T., Hua, W., Georg, M., Grundmann, M., 2019. MediaPipe: A Framework for Building Perception Pipelines.

Moetia Putri, H., Fuadi, W., 2020. Pendeteksian Bahasa Isyarat Indonesia Secara Real-time menggunakan Long Short Term Memory (LSTM). Tts 1, 1–13.

Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, Avidan, S., 2006. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06). Presented at the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06), IEEE, New York, NY, USA, pp. 1491–1498. https://doi.org/10.1109/CVPR.2006.119

Shamitha S H, S.H., Badarinath K, K., 2023. Sign Language Recognition utilizing LSTM and Mediapipe for Dynamic Gestures of ISL. Int. J. Multidiscip. Res. 5, 6868. https://doi.org/10.36948/ijfmr.2023.v05i05.6868