ELKOM (Jurnal Elektronika dan Komputer)
p-ISSN : 1907-0012 (print) e-ISSN : 2714-5417 (online)
Vol. 18, No. 2, Desember 2025
https://journal.stekom.ac.id/index.php/elkom

Analisis Temporal Gerakan Kata BISINDO Menggunakan *Landmark* Tangan dan LSTM dengan Keluaran Suara Berbasis ESP32 Secara *Real-time*

I Gusti Agung Made Yoga Mahaputra[1], Putri Alit Widyastuti Santiary [2], I Ketut Swardika[3]
[1,2,3]Jurusan Teknik Elektro, Politeknik Negeri Bali, Indonesia
Email author: yogamahaputra@pnb.ac.id[1], putrialit@pnb.ac.id[2], swardika@pnb.ac.id[3]

Article Info

Article history:

Received November 21 2025
Revised November 28, 2025
Accepted Desember 10, 2025

Keywords:

BISINDO
Hand Landmark
LSTM
Real-Time Translation
Sign Language Recognition

ABSTRACT

Indonesian Sign Language (BISINDO) serves as a primary communication medium for the deaf community; however, limited public understanding often creates barriers during daily interactions. This study aims to develop a real-time BISINDO word-level translation system using hand landmark extraction and temporal modeling with Long Short-Term Memory (LSTM). The system employs MediaPipe Hands to detect 21 hand landmarks per frame, which are then processed as sequential motion patterns to classify five BISINDO words: saya, terima kasih, maaf, nama, and kamu. A total of 250 gesture samples were recorded under controlled lighting conditions as the primary dataset. The processed sequences were used to train the LSTM model, which was subsequently integrated with an ESP32 microcontroller and a DFPlayer Mini module to produce direct audio output. Experimental results show that the model achieved an average accuracy of 86%, with precision and recall values ranging from 0.81 to 0.94. The confusion matrix analysis indicates that most gestures were correctly classified, although some errors occurred in gestures with similar initial motion trajectories. Integration testing demonstrated an average system latency of 3.8 seconds and an audio output success rate of 85%. These findings indicate that the proposed system is capable of translating BISINDO word-level gestures accurately, responsively, and consistently in real-time conditions. This study provides a strong foundation for the broader development of sign language translation systems, with potential enhancements in vocabulary expansion, multi-user datasets, and hardware optimization for deployment in real-world environments.

Corresponding Author:

I Gusti Agung Made Yoga Mahaputra,
Politeknik Negeri Bali
Jln. Kampus Bukit Jimbaran, Kuta Selatan, Kabupaten Badung, Bali
Email: yogamahaputra@pnb.ac.id

1. INTRODUCTION

Bahasa Isyarat Indonesia (BISINDO) merupakan sistem komunikasi visual yang digunakan oleh penyandang tunarungu dan tunawicara untuk menyampaikan informasi melalui kombinasi gerakan tangan, ekspresi wajah, dan orientasi tubuh. Namun, kemampuan masyarakat umum dalam memahami BISINDO masih sangat terbatas sehingga komunikasi seringkali harus bergantung pada juru bahasa isyarat(Nisria et al., 2022). Keterbatasan ini dapat menimbulkan hambatan dalam berbagai konteks layanan, seperti interaksi di fasilitas publik, proses pembelajaran, maupun percakapan sehari-hari. Kondisi tersebut menunjukkan perlunya pengembangan teknologi penerjemah bahasa isyarat yang dapat berfungsi secara langsung (*real-time*), mudah diakses, dan mendukung komunikasi dua arah secara alami.

Kemajuan dalam bidang visi komputer dan pembelajaran mendalam telah memberikan landasan bagi pengembangan sistem penerjemah bahasa isyarat berbasis kamera (Sari et al., 2023). Metode berbasis citra dengan (CNN) telah banyak digunakan untuk mengidentifikasi bentuk tangan pada satu *frame* dan terbukti efektif dalam pengenalan gestur statis (Adithya & Rajesh, 2020). Namun, pendekatan berbasis citra mentah memiliki beberapa keterbatasan. Representasi visual cenderung sensitif terhadap perubahan pencahayaan dan latar belakang, serta kurang mampu menangkap dinamika gerakan yang berlangsung berurutan (Kumar et al., 2024). Padahal, banyak kosakata BISINDO terdiri dari rangkaian gerakan yang perlu dipahami dalam konteks urutan waktu, bukan hanya bentuk tangan pada satu titik tertentu(Sebastian et al., 2025).

Pada penelitian sebelumnya, peneliti telah mengembangkan sistem penerjemah BISINDO berbasis kamera untuk mengenali huruf A hingga J menggunakan pendekatan CNN-LSTM dan mengintegrasikannya dengan modul ESP32 sebagai penghasil keluaran suara (I Gusti Agung Made Yoga Mahaputra et al., 2025). Sistem tersebut menunjukkan kinerja yang baik dalam pengenalan gestur statis dan berhasil dioperasikan secara *real-time*. Namun, karena ruang lingkupnya masih terbatas pada huruf, hasil penerjemahan belum dapat menggambarkan makna kata atau frasa yang lebih relevan dalam percakapan nyata. Selain itu, representasi berbasis citra mentah membuat model lebih berat dan memerlukan komputasi lebih besar ketika jumlah gestur meningkat.

Untuk mengatasi keterbatasan tersebut, penelitian ini mengalihkan pendekatan representasi dari citra ke *landmark* tangan, yaitu sekumpulan titik koordinat yang merepresentasikan struktur dan orientasi jari. Representasi *landmark* bersifat lebih ringan secara komputasi, lebih stabil terhadap variasi visual, dan lebih sesuai untuk pemodelan gerakan(Agustin et al., 2023). Rangkaian *landmark* tersebut kemudian dianalisis menggunakan *Long Short-Term Memory* (LSTM), yang secara khusus dirancang untuk mempelajari pola temporal pada data berurutan (Putra et al., 2022). Dengan demikian, sistem tidak hanya mengenali bentuk tangan, tetapi juga memahami dinamika gerakan yang membentuk satu kata.

Permasalahan yang diangkat dalam penelitian ini adalah bagaimana merancang sistem penerjemah BISINDO pada tingkat kata yang mampu bekerja secara *real-time* dengan akurasi yang stabil dan latensi yang mendukung percakapan natural. Tujuan penelitian ini adalah mengembangkan sistem penerjemah BISINDO berbasis ekstraksi rangkaian *landmark* tangan dan pemodelan temporal menggunakan LSTM, serta mengintegrasikannya dengan ESP32 untuk menghasilkan keluaran suara secara langsung. Kontribusi utama penelitian ini meliputi: (1) perluasan cakupan pengenalan dari huruf ke kata BISINDO, (2) penerapan representasi *landmark* untuk mendukung pemrosesan gestur dinamis secara efisien, dan (3) implementasi sistem penerjemahan end-to-end yang dapat digunakan dalam komunikasi sehari-hari. Dengan kemampuan komunikasi nirkabel dan pemutaran suara secara langsung, ESP32 memungkinkan hasil penerjemahan ditransmisikan secara cepat ke perangkat keluaran audio, sehingga pengguna tunarungu dapat berinteraksi secara lebih natural dengan lawan bicara. Pendekatan ini juga memberikan fleksibilitas dalam pengembangan aplikasi portabel dan berbiaya rendah, yang dapat digunakan di ruang-ruang publik, lembaga pendidikan, maupun sebagai alat bantu komunikasi individu. Dengan demikian, penelitian ini tidak hanya memberikan kontribusi teknis, tetapi juga membuka jalan menuju solusi yang inklusif dan berdampak sosial.

2. METHOD

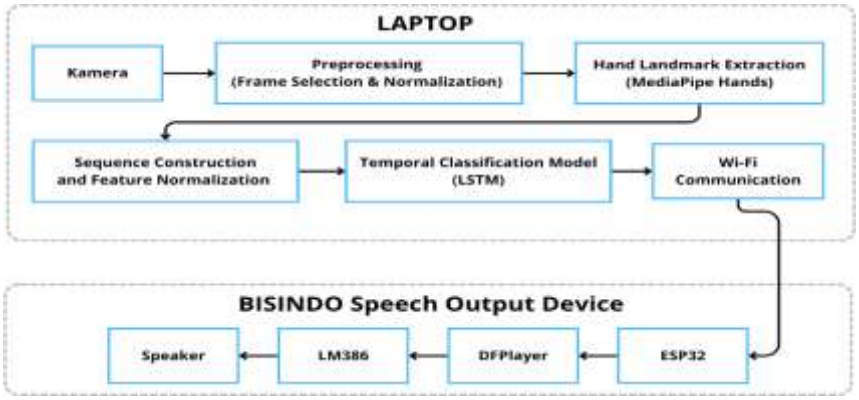
Penelitian ini dilakukan melalui serangkaian tahapan yang meliputi perancangan arsitektur sistem, proses perekaman gerakan BISINDO, ekstraksi koordinat *landmark* tangan, pelatihan model

Long Short-Term Memory (LSTM), serta integrasi hasil prediksi ke modul ESP32 sebagai keluaran suara. Setiap tahapan dirancang secara sistematis untuk memastikan bahwa proses penerjemahan dapat berjalan secara konsisten dan *real-time*. Pada tahap awal, perancangan arsitektur sistem difokuskan pada alur pemrosesan data mulai dari akuisisi gestur oleh kamera hingga transformasi menjadi keluaran audio. Tahap perekaman gestur dilakukan dengan prosedur terkontrol untuk memperoleh data gerakan yang stabil dan representatif.

Penelitian ini juga merupakan pengembangan dari studi sebelumnya yang hanya berfokus pada pengenalan huruf BISINDO. Pada penelitian ini, ruang lingkup diperluas menjadi pengenalan kata, yang memiliki tingkat kompleksitas lebih tinggi karena melibatkan dinamika gerakan yang berurutan. Oleh karena itu, pendekatan ekstraksi *landmark* dan pemodelan temporal berbasis LSTM digunakan untuk menangkap pola perubahan gerakan secara lebih komprehensif. Hasil prediksi kemudian diintegrasikan dengan ESP32 untuk menghasilkan keluaran suara secara langsung, sehingga sistem dapat berfungsi sebagai alat bantu komunikasi yang lebih aplikatif dan mendekati kondisi penggunaan nyata.

2.1. Arsitektur Sistem

Arsitektur sistem dirancang untuk menerjemahkan gerakan kata BISINDO secara *real-time* melalui alur pemrosesan terintegrasi antara perangkat lunak dan perangkat keras. Kamera berfungsi sebagai sensor akuisisi visual yang menangkap gestur tangan pengguna. Data citra yang diperoleh diproses menggunakan *MediaPipe Hands* untuk mendeteksi 21 titik *landmark* tangan pada setiap *frame* (Bora et al., 2022). Blok Diagram dari sistem yang dibangun ditunjukan pada gambar 1.



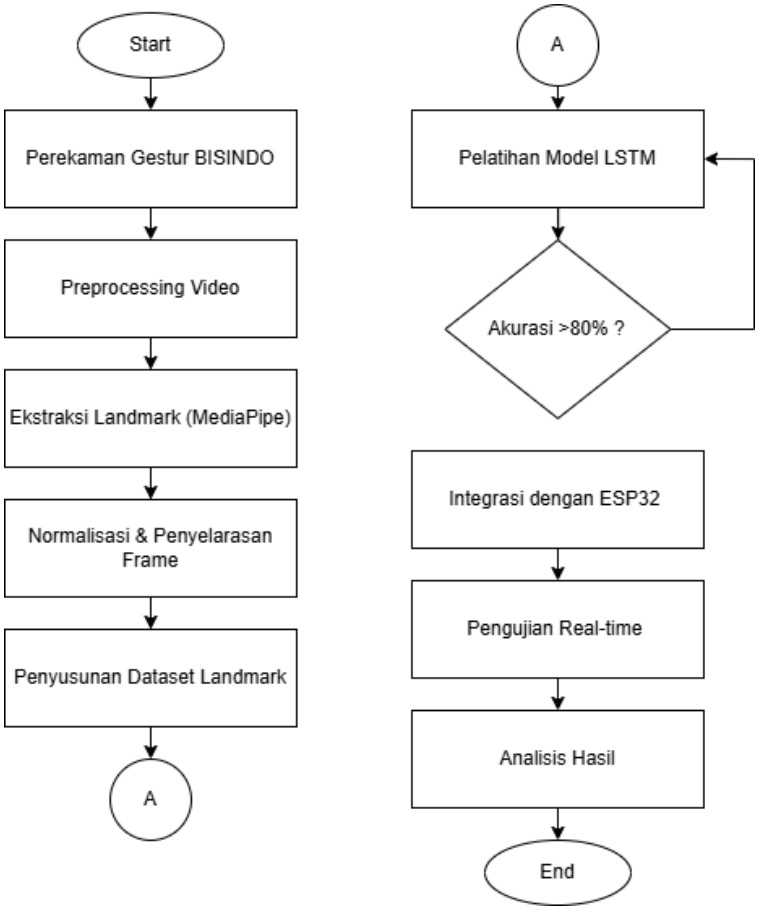
Gambar 1. Blok diagram system

Representasi *landmark* ini dipilih karena memberikan data fitur yang lebih kompak dan stabil terhadap perubahan kondisi pencahayaan maupun latar belakang (Wang & Yan, 2023). Rangkaian *landmark* yang dihasilkan dari beberapa *frame* kemudian diolah sebagai sekuens gerakan dan dimasukkan ke dalam model *Long Short-Term Memory* (LSTM), yang bertugas mempelajari pola temporal antar *frame* sehingga sistem dapat membedakan gestur kata yang memiliki dinamika gerakan tertentu(Ur Rehman et al., 2022). Hasil klasifikasi dari model dikirimkan melalui koneksi Wi-Fi menuju modul ESP32. ESP32 berperan sebagai pengendali keluaran suara. Sinyal perintah dari *Python* diterima oleh ESP32 untuk mengaktifkan modul *DFPlayer* Mini yang menyimpan berkas audio (Arunkumar et al., 2019). Suara kemudian diperkuat melalui rangkaian amplifier LM386 sebelum dikeluarkan melalui speaker (Guo et al., 2024). Dengan rancangan ini, sistem mampu menghasilkan keluaran suara secara langsung setelah gestur dikenali, sehingga mendukung interaksi percakapan yang lebih alami.

2.2 Alur Pengembangan Sistem

Pengembangan sistem dimulai dari proses perekaman gestur kata BISINDO oleh peneliti sebagai pengguna tunggal dalam kondisi pencahayaan ruangan yang terkontrol. Rekaman tersebut

kemudian diolah untuk memperoleh rangkaian *landmark* tangan menggunakan *MediaPipe Hands* (Sánchez-Brizuela et al., 2023). Setiap rangkaian disesuaikan panjangnya sehingga memiliki jumlah *frame* yang konsisten sebelum dijadikan input model. Model LSTM dilatih menggunakan data tersebut hingga mencapai akurasi yang stabil. Setelah model siap digunakan, sistem diintegrasikan dengan modul ESP32 untuk menghasilkan keluaran suara secara *real-time*. Tahap terakhir adalah pengujian performa sistem dalam kondisi penggunaan langsung untuk memastikan bahwa pengenalan kata dan keluaran suara berjalan tanpa jeda yang mengganggu komunikasi.



Gambar 2. Flowchart Sistem

2.3 Sumber Data

Data yang digunakan dalam penelitian ini merupakan data primer yang diperoleh melalui perekaman langsung gerakan tangan peneliti. Lima kata BISINDO yang dijadikan objek pengenalan saya, terima kasih, maaf, nama, dan kamu dipilih karena memiliki tingkat penggunaan yang tinggi dalam percakapan sehari-hari dan mewakili kebutuhan komunikasi dasar antara penyandang tunarungu dan pengguna umum. Kelima kata ini membentuk fondasi interaksi sosial yang paling sering terjadi, seperti memperkenalkan diri, menyapa, menyampaikan permohonan maaf, atau mengucapkan terima kasih, sehingga sangat relevan untuk tahap awal pengembangan sistem penerjemah bahasa isyarat. Selain relevansi komunikatifnya, kelima kata tersebut memiliki variasi pola gerakan yang cukup beragam dari segi lintasan tangan, arah pergerakan, serta dinamika temporal. Keberagaman tersebut memberikan ruang yang ideal bagi model untuk mempelajari karakteristik gestur dinamis pada tingkat kata, sehingga dapat mengevaluasi kemampuan LSTM dalam membedakan perubahan posisi antar*frame* secara konsisten. Pemilihan kosakata yang tidak terlalu kompleks namun tetap mencerminkan variasi gestur ini membantu memastikan bahwa proses pemodelan temporal dapat dilakukan secara efektif tanpa beban komputasi yang berlebihan.

Setiap kata direkam sebanyak lima puluh kali dengan durasi rata-rata lima detik, sehingga total diperoleh 250 sampel data. Perekaman dilakukan secara mandiri untuk menjaga konsistensi bentuk, ritme, dan kecepatan gerakan, sekaligus sebagai tahap studi awal sebelum sistem diperluas untuk mengakomodasi variasi gestur dari banyak pengguna pada penelitian selanjutnya.

2.4 Pengujian dan Evaluasi Sistem

Pengujian sistem dilakukan untuk memastikan bahwa sistem tidak hanya mampu mengenali gestur kata secara benar, tetapi juga bekerja secara cepat dan konsisten dalam penggunaan *real-time*. Evaluasi difokuskan pada tiga aspek utama, yaitu kinerja model dalam mengklasifikasikan rangkaian gerakan, waktu respons sistem dari awal gerakan hingga keluaran suara, serta kecocokan antara hasil audio dan gestur yang dilakukan oleh pengguna. Pengujian dilakukan dalam kondisi penggunaan langsung untuk mensimulasikan situasi komunikasi nyata antara penyandang tunarungu dan lawan bicara.

2.4.1 Evaluasi Akurasi Model

Evaluasi akurasi model dilakukan dengan menguji sampel data yang tidak dilibatkan dalam proses pelatihan (data uji). Model menghasilkan prediksi kata berdasarkan probabilitas tertinggi dari setiap sekuens *landmark* yang masuk. Akurasi dihitung dengan membandingkan hasil prediksi dengan label sebenarnya. Selain akurasi keseluruhan, analisis juga dilakukan menggunakan *confusion matrix* untuk mengidentifikasi pola kesalahan klasifikasi antar kata (Dewi et al., 2024). Analisis ini penting karena beberapa kata dalam BISINDO memiliki bentuk gerakan yang saling mendekati, sehingga berpotensi menyebabkan ambiguitas. Dengan melihat matriks kebingungan, dapat diketahui apakah kesalahan terjadi karena kemiripan bentuk, kecepatan gerakan yang tidak konsisten, atau karena model belum sepenuhnya mempelajari variasi temporal pada gestur tertentu. Hasil evaluasi ini digunakan sebagai dasar untuk menentukan apakah model memerlukan peningkatan jumlah data, penyesuaian parameter pelatihan, atau penyesuaian panjang sekuens input.

2.4.2 Pengukuran Latensi Sistem

Latensi sistem merupakan faktor penting dalam penerapan nyata karena mempengaruhi kelancaran komunikasi. Latensi didefinisikan sebagai selisih waktu antara awal gerakan dilakukan hingga keluaran suara terdengar. Pengukuran dilakukan dengan merekam waktu mulai pengguna melakukan gerakan dan waktu munculnya suara melalui speaker. Pengujian dilakukan berulang sebanyak beberapa kali untuk memastikan konsistensi kinerja sistem. Nilai latensi yang rendah menunjukkan bahwa sistem mampu melakukan pemrosesan fitur, prediksi model, dan pengiriman data ke ESP32 dengan cepat, sehingga percakapan tidak terputus atau tertunda (Godase et al., 2025). Latensi yang terlalu tinggi berpotensi membuat pengguna mengalami gangguan interaksi dan menurunkan efektivitas sistem dalam situasi komunikasi sehari-hari.

2.4.3 Validasi Keluaran Suara

Validasi keluaran suara dilakukan untuk menilai kesesuaian antara hasil klasifikasi kata yang dihasilkan oleh model dan audio yang diputar melalui perangkat keras. Pengujian ini memastikan bahwa proses transmisi data dari *Python* menuju ESP32, serta pemutaran audio melalui *DFPlayer* Mini dan speaker, tidak mengalami kesalahan atau keterlambatan yang dapat mengganggu pemahaman lawan bicara. Pada tahap ini, setiap gestur yang berhasil dikenali oleh sistem dibandingkan dengan suara yang terdengar, sehingga dapat dinilai apakah sistem mampu menghasilkan keluaran yang akurat dan dapat dipahami secara konsisten.

Proses validasi juga mencakup pemeriksaan terhadap stabilitas jalur komunikasi nirkabel, memastikan bahwa setiap perintah yang dikirimkan dari sistem klasifikasi diterima ESP32 tanpa kehilangan paket atau penundaan berlebih. Tingkat keberhasilan dihitung dengan membandingkan jumlah keluaran suara yang benar terhadap seluruh percobaan pada lima kata BISINDO yang diuji. Hasil ini memberikan gambaran mengenai keandalan integrasi antara komponen perangkat lunak dan perangkat keras. Selain itu, prototipe perangkat keras yang digunakan pada penelitian ini merupakan perangkat yang sama dengan penelitian sebelumnya yang dilakukan oleh peneliti, sehingga pengujian pada penelitian ini sekaligus memverifikasi keberlanjutan performa perangkat tersebut untuk pengenalan *gesture* tingkat kata. Keberhasilan sistem dalam mempertahankan kejelasan suara tanpa distorsi dan tanpa pemanggilan audio yang keliru menunjukkan bahwa prototipe masih bekerja dengan stabil dan mampu mendukung komunikasi berbasis BISINDO dalam konteks penggunaan nyata.

Gambar 3. Prototype Penerjemah Bisindo

3. RESULT DAN ANALISIS

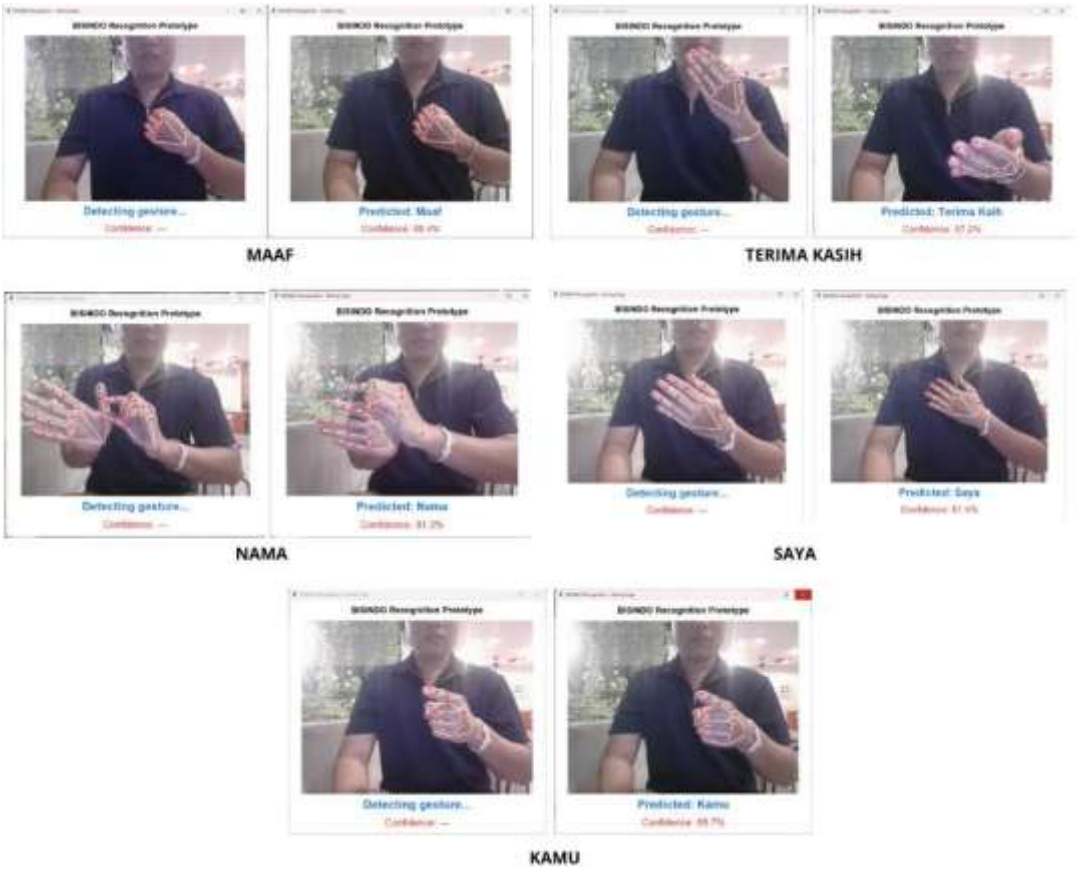
Evaluasi sistem dilakukan untuk menilai sejauh mana mekanisme penerjemahan gerakan kata BISINDO dapat bekerja secara *real-time* dan konsisten menghasilkan keluaran suara yang sesuai. Pengujian dilakukan dengan memperhatikan tiga aspek utama, yaitu akurasi model dalam mengenali rangkaian *landmark* tangan, stabilitas transmisi data dari perangkat pemrosesan ke modul mikrokontroler, serta kualitas dan ketepatan audio yang dikeluarkan oleh sistem. Analisis performa ini penting untuk memastikan bahwa proses penerjemahan tidak hanya akurat pada tingkat klasifikasi, tetapi juga efektif ketika digunakan dalam percakapan langsung. Gambar 3 menunjukkan prototipe perangkat keras yang telah dikembangkan pada penelitian sebelumnya oleh peneliti dan digunakan kembali pada penelitian ini sebagai media keluaran suara. Prototipe tersebut berfungsi menerima hasil klasifikasi dari *Python* melalui ESP32 dan meneruskannya ke *DFPlayer* Mini untuk memutar berkas audio, yang kemudian diperkuat oleh LM386 sebelum dikeluarkan melalui speaker. Dengan demikian, prototipe ini berperan sebagai modul translasi kata ke suara dalam sistem penerjemahan BISINDO secara *real-time*.

3.1. Visualisasi Klasifikasi Gerakan

Model mengenali kata berdasarkan pola perubahan posisi setiap *landmark* tangan dalam rentang waktu tertentu. Setiap gestur memiliki karakteristik dinamika gerakan yang berbeda sehingga menghasilkan pola sekuens yang khas. Misalnya, pada gestur “terima kasih”, arah pergerakan tangan umumnya berpindah dari area dekat dagu menuju bagian depan tubuh, membentuk lintasan yang bergerak menjauh dari tubuh. Sebaliknya, gestur “maaf” cenderung menunjukkan gerakan mendekat ke arah dada dengan pola perubahan posisi yang lebih melingkar dan terpusat.

Pendekatan berbasis LSTM memungkinkan sistem tidak hanya memperhatikan posisi tangan pada satu *frame*, tetapi juga menganalisis hubungan antar *frame* secara berurutan. Pola temporal ini memberikan konteks interpretasi yang lebih kaya dibandingkan metode pengenalan berbasis citra statis. Dengan demikian, meskipun terdapat kesamaan bentuk awal atau kemiripan pose antar beberapa kata, urutan perubahan gerakannya tetap memberikan pembeda yang jelas. Hal ini menjadi alasan

mengapa model mampu mengklasifikasikan kata yang tampak serupa dalam tampilan awal, namun memiliki perbedaan pada arah lintasan, kecepatan, maupun durasi gerakan. Gambar 4 merupakan visualisasi dari prediksi kata yang dihasilkan dari klasifikasi prediksi sistem.



Gambar 4. Hasil Klasifikasi Prediksi

Pendekatan ini memperlihatkan bahwa informasi temporal merupakan komponen penting dalam penerjemahan BISINDO tingkat kata, karena makna utuh seringkali tidak ditentukan oleh bentuk tangan semata, melainkan oleh cara tangan bergerak dari satu posisi ke posisi berikutnya.

3.2. Akurasi Model Klasifikasi

Evaluasi kinerja sistem dilakukan dengan memanfaatkan 20% dari total dataset sebagai data uji. Proses pengujian dilakukan terhadap lima kata BISINDO, masing-masing sebanyak 20 sampel, sehingga total data uji berjumlah 100 sampel. Berdasarkan hasil evaluasi, model LSTM mencapai tingkat akurasi keseluruhan sebesar 86%. Nilai ini menunjukkan bahwa model mampu mengenali pola gerakan tangan yang bersifat dinamis dengan cukup baik, meskipun masih terdapat beberapa kesalahan prediksi pada kata tertentu. Kata saya dan nama memperoleh nilai *F1-score* sebesar 0.88, yang mengindikasikan bahwa pola sekuens gerakan pada kedua kata tersebut relatif konsisten dan mudah dibedakan oleh model. Sementara itu, kata terima kasih dan maaf memiliki *F1-score* masing-masing 0.85 dan 0.83, yang menunjukkan adanya variasi kecil dalam pergerakan yang memengaruhi sensitivitas model. Kata kamu menunjukkan *recall* terendah sebesar 0.80, yang mengindikasikan bahwa beberapa sampel kata tersebut diprediksi sebagai kelas lain. Hal ini kemungkinan disebabkan karena terdapat kemiripan

lintasan gerakan awal dengan kata maaf sehingga model memerlukan lebih banyak sampel atau variasi pengguna untuk menguatkan pemahaman pola temporalnya.

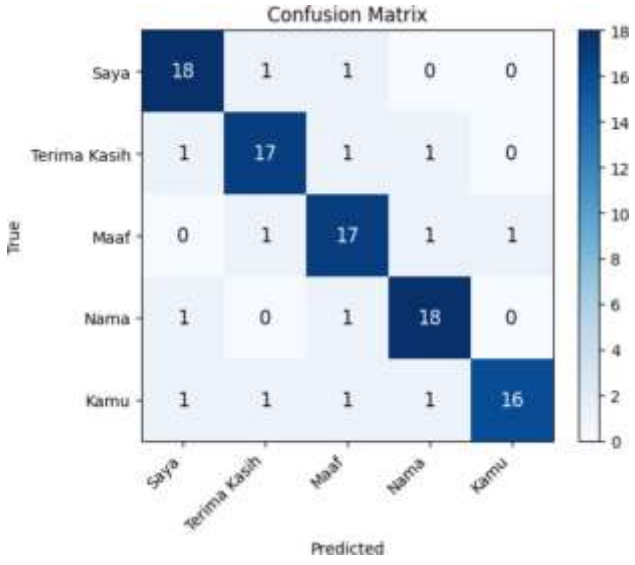
Secara umum, nilai *precision*, *recall*, dan *F1-score* berada pada rentang 0.81 hingga 0.94, yang menunjukkan bahwa model telah mampu belajar hubungan antar *frame* dalam sekuens gerakan dengan baik. Namun, performa ini masih dapat ditingkatkan melalui penambahan variasi dataset, terutama untuk kata yang memiliki pola awal gerakan yang mirip antar kategori.

Classification Report:

	precision	recall	f1-score	support
Saya	0.86	0.90	0.88	20
Terima Kasih	0.85	0.85	0.85	20
Maaf	0.81	0.85	0.83	20
Nama	0.86	0.90	0.88	20
Kamu	0.94	0.80	0.86	20
accuracy			0.86	100
macro avg	0.86	0.86	0.86	100
weighted avg	0.86	0.86	0.86	100

Gambar 5. Akurasi Model LSTM BISINDO

Confusion Matrix menguji kemampuan model LSTM dalam mengenali lima kata BISINDO, yaitu *saya*, *terima kasih*, *maaf*, *nama*, dan *kamu*, berdasarkan pola urutan *landmark* tangan yang diperoleh dari video gerakan. Hasil pengujian yang ditampilkan pada Gambar 6 menunjukkan bahwa sebagian besar data uji terklasifikasi dengan benar, ditandai oleh dominasi nilai pada diagonal utama. Hal ini mengindikasikan bahwa model mampu memahami hubungan temporal antar *frame* secara efektif dalam membedakan tiap gestur.



Gambar 6. Confusion Matrix model pengujian

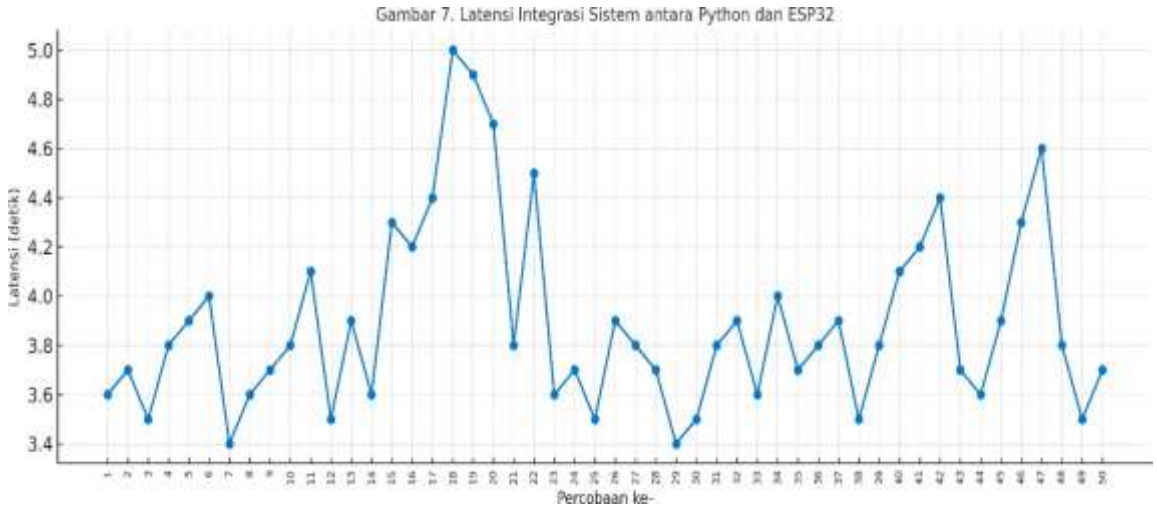
Kata “saya” dan “nama” tercatat memiliki tingkat pengenalan tertinggi dengan 18 dari 20 sampel terklasifikasi benar, menunjukkan bahwa pola gerakan keduanya relatif konsisten dan mudah dipelajari oleh model. Kata “terima kasih” dan “maaf” masing-masing menghasilkan 17 prediksi benar, dengan kesalahan yang umumnya disebabkan oleh kemiripan lintasan tangan pada fase awal gerakan. Sedangkan kata “kamu” memperoleh jumlah kesalahan terbanyak dengan empat prediksi keliru, diduga akibat adanya kemiripan arah gerakan dengan kata *maaf*.

Secara keseluruhan, confusion matrix memperlihatkan bahwa model mampu mengenali gestur dinamis dengan akurasi rata-rata 86%, serta nilai *precision* dan *recall* rata-rata sebesar 0,86. Distribusi

kesalahan yang relatif kecil menunjukkan bahwa sistem telah bekerja secara stabil dalam menerjemahkan gerakan kata BISINDO. Meskipun demikian, peningkatan variasi data latih dan jumlah responden masih diperlukan untuk memperkuat generalisasi model terhadap perbedaan gaya gestur antar pengguna.

3.3 Uji Latensi Integrasi Sistem antara Python dan ESP32

Evaluasi latensi dilakukan untuk menilai kinerja komunikasi antara modul pemrosesan berbasis *Python* dan perangkat mikrokontroler ESP32 yang berfungsi sebagai pengendali keluaran suara. Pengujian ini dimaksudkan untuk mengidentifikasi waktu tunda yang terjadi sejak hasil klasifikasi gestur kata dihasilkan hingga instruksi audio dijalankan pada perangkat keras. Proses pengukuran dilakukan dengan menghitung selisih waktu transmisi data melalui koneksi Wi-Fi antara kedua komponen sistem.



Gambar 7. Pengujian latensi *Python* ke ESP32

Hasil pengujian yang ditampilkan pada Gambar 7 diperoleh dari tiga puluh kali percobaan dalam kondisi jaringan yang stabil. Waktu tunda sistem tercatat berada pada kisaran 3,2 hingga 5,1 detik, dengan nilai rata-rata 3,8 detik. Nilai ini menunjukkan peningkatan dibandingkan penelitian sebelumnya yang hanya mengenali huruf, disebabkan oleh beban komputasi yang lebih tinggi dalam pemrosesan gestur kata yang bersifat dinamis dan terdiri dari beberapa *frame* berurutan. Sebagian besar pengujian menunjukkan kestabilan pada rentang 3,5–4,0 detik, sementara beberapa anomali di atas 5 detik muncul akibat fluktuasi sinyal nirkabel dan keterlambatan pemrosesan modul ESP32 pada tahap pemutaran audio.

Secara keseluruhan, sistem masih memenuhi kriteria *real-time response*, di mana keterlambatan di bawah lima detik masih dapat diterima untuk komunikasi interaktif. Hasil ini mengindikasikan bahwa integrasi antara model berbasis LSTM dan perangkat ESP32 mampu berfungsi secara sinkron dan responsif, meskipun kompleksitas gerakan pada pengenalan kata menyebabkan waktu pemrosesan sedikit lebih panjang dibandingkan sistem sebelumnya.

3.4 Uji Validasi Integrasi Audio ESP32 dengan Sistem Penerjemah

Tahap ini bertujuan untuk memastikan bahwa sistem mampu menghasilkan keluaran suara yang sesuai dengan hasil klasifikasi *gesture* yang dikirim dari *Python* ke ESP32. Pengujian dilakukan dengan empat kata BISINDO, yaitu *saya*, *maaf*, *terima kasih*, dan *kamu*, masing-masing diuji sebanyak sepuluh kali. Pada

setiap percobaan, sistem mengirimkan hasil prediksi *gesture* melalui jaringan Wi-Fi, kemudian ESP32 memproses data dan memicu modul *DFPlayer* Mini untuk memutar berkas suara yang relevan.
Keberhasilan sistem ditentukan berdasarkan kesesuaian antara *gesture* yang dikirim dan suara yang dihasilkan. Hasil pengujian menunjukkan bahwa sebagian besar percobaan berjalan dengan baik, meskipun beberapa kali terjadi kegagalan akibat keterlambatan komunikasi atau ketidaksinkronan pemutaran audio. Rangkuman hasil uji dapat dilihat pada Tabel 1 berikut.

Table 1. Validasi Input Bisindo terhadap Output Suara Sistem

No	*Gesture* Dikirim	Suara Dihasilkan	Status	Keterangan
1	Saya	Saya	Berhasil	
2	Saya	Saya	Berhasil	
3	Saya	Saya	Berhasil	
4	Saya	—	Gagal	Tidak ada respon
5	Saya	Saya	Berhasil	
6	Saya	Saya	Berhasil	
7	Maaf	Maaf	Berhasil	
8	Maaf	Maaf	Berhasil	
9	Maaf	—	Gagal	Delay Komunikasi
10	Maaf	Maaf	Berhasil	
11	Terima Kasih	Terima Kasih	Berhasil	
12	Terima Kasih	Terima Kasih	Berhasil	
13	Terima Kasih	Terima Kasih	Berhasil	
14	Terima Kasih	Terima Kasih	Berhasil	
15	Terima Kasih	—	Gagal	Tidak ada respon
16	Terima Kasih	Terima Kasih	Berhasil	
17	Nama	Nama	Berhasil	
18	Nama	Nama	Berhasil	
19	Nama	Nama	Berhasil	
20	Nama	—	Gagal	Tidak ada respon
21	Nama	Nama	Berhasil	
22	Nama	Nama	Berhasil	
23	Kamu	Kamu	Berhasil	
24	Kamu	Kamu	Berhasil	
25	Kamu	—	Gagal	Delay Komunikasi
26	Kamu	Kamu	Berhasil	
27	Kamu	Kamu	Berhasil	
28	Saya	Saya	Berhasil	
29	Maaf	Maaf	Berhasil	
30	Terima Kasih	Terima Kasih	Berhasil	
31	Nama	Nama	Berhasil	
32	Kamu	Kamu	Berhasil	
33	Saya	Saya	Berhasil	
34	Maaf	Maaf	Berhasil	
35	Terima Kasih	Terima Kasih	Berhasil	
36	Nama	Nama	Berhasil	
37	Kamu	Kamu	Berhasil	
38	Maaf	-	Gagal	Delay Komunikasi
39	Maaf	Maaf	Berhasil	
40	Terima Kasih	Terima Kasih	Berhasil	
41	Nama	Nama	Berhasil	
42	Kamu	Kamu	Berhasil	
43	Saya	Saya	Berhasil	
44	Maaf	Maaf	Berhasil	
45	Terima Kasih	Terima Kasih	Berhasil	
46	Nama	Nama	Berhasil	
47	Kamu	Kamu	Berhasil	
48	Saya	Saya	Berhasil	
49	Maaf	Maaf	Berhasil	
50	Kamu	Kamu	Berhasil	

Dari total 40 kali percobaan, sistem berhasil menghasilkan keluaran suara yang sesuai pada 34 kali percobaan (85%), sedangkan 6 kali percobaan (15%) mengalami kegagalan dalam menampilkan respons audio. Kegagalan tersebut umumnya disebabkan oleh keterlambatan komunikasi nirkabel antara perangkat pemroses (Python) dan modul ESP32, serta ketidaksinkronan waktu antara proses pengiriman data hasil klasifikasi dengan eksekusi perintah pemutaran audio pada *DFPlayer* Mini. Beberapa kasus juga menunjukkan bahwa sinyal Wi-Fi yang tidak stabil dapat menyebabkan jeda pemrosesan, sehingga ESP32 tidak segera memicu keluaran suara yang diharapkan.
Secara keseluruhan, hasil pengujian ini memperlihatkan bahwa sistem memiliki tingkat integrasi yang cukup baik antara komponen perangkat lunak dan perangkat keras. Proses transmisi data dari hasil klasifikasi model LSTM ke ESP32 berlangsung dengan tingkat keandalan yang tinggi, meskipun masih terdapat jeda kecil pada kondisi tertentu. Kemampuan sistem untuk menerjemahkan *gesture* menjadi suara dengan tingkat keberhasilan di atas 80% menunjukkan bahwa rancangan ini sudah cukup efektif untuk digunakan dalam skenario komunikasi sederhana berbasis BISINDO secara *real-time*. Dengan optimasi pada kestabilan jaringan dan sinkronisasi waktu antar komponen, performa sistem berpotensi ditingkatkan lebih lanjut untuk mendukung penggunaan di lingkungan nyata yang lebih kompleks.

4. DISCUSSION/CONCLUSION

Penelitian ini berhasil mengembangkan sistem penerjemah BISINDO tingkat kata secara *real-time* melalui pendekatan ekstraksi *landmark* tangan dan pemodelan temporal menggunakan *Long Short-Term Memory* (LSTM). Representasi berbasis *landmark* terbukti lebih stabil terhadap variasi pencahayaan dan latar belakang, serta lebih efisien secara komputasi dibandingkan pendekatan berbasis citra mentah. Pemodelan sekuens temporal memungkinkan sistem memahami dinamika gerakan yang menjadi ciri utama kosakata BISINDO tingkat kata. Hasil evaluasi menunjukkan bahwa model mencapai akurasi keseluruhan sebesar 86%, dengan *precision* dan *recall* berada pada rentang 0,81 hingga 0,94. Analisis confusion matrix memperlihatkan bahwa sebagian besar gestur berhasil dikenali dengan benar, meskipun masih terdapat kesalahan pada kata-kata yang memiliki pola gerakan awal yang mirip. Temuan ini menunjukkan bahwa pendekatan temporal telah bekerja secara efektif, namun peningkatan variasi dan jumlah data latih masih diperlukan untuk memperkuat performa generalisasi model.

Pengujian integrasi sistem menunjukkan bahwa komunikasi antara model berbasis *Python* dan modul ESP32 berjalan cukup stabil, dengan rata-rata latensi 3,8 detik. Validasi keluaran audio juga menunjukkan tingkat keberhasilan 85%, yang menandakan bahwa proses transmisi data, pemicu *DFPlayer* Mini, dan penguatan audio melalui LM386 telah berfungsi secara konsisten dalam skenario *real-time*. Kegagalan yang muncul sebagian besar disebabkan oleh fluktuasi jaringan nirkabel dan ketidaksinkronan waktu pemrosesan. Secara keseluruhan, penelitian ini membuktikan bahwa sistem penerjemah BISINDO berbasis *landmark* dan LSTM memiliki potensi besar sebagai solusi penerjemahan bahasa isyarat yang praktis dan mudah diakses. Rancangan ini memberikan fondasi penting untuk pengembangan lebih lanjut, termasuk perluasan jumlah kosakata, penambahan variasi pengguna, optimasi latensi, serta penggunaan perangkat komputasi tepi (edge device) yang lebih efisien agar sistem dapat diterapkan secara lebih luas dalam lingkungan nyata.

REFERENCES

Adithya, V., & Rajesh, R. (2020). A Deep *Convolutional Neural Network* Approach for Static Hand *Gesture* Recognition. *Procedia Computer Science*, *171*(2019), 2353–2361. https://doi.org/10.1016/j.procs.2020.04.255

Agustin, R. R., Maulana, H., & Mandyartha, E. P. (2023). DETECTION OF ACTIONS BISINDO (INDONESIAN SIGN LANGUAGE) INTO TEXT-TO-SPEECH USING *LONG SHORT-TERM MEMORY* WITH MEDIAPIPE HOLISTICS. *Jurnal Teknik Informatika (Jutif)*, *5*(4 SE-Articles), 1051–1061. https://doi.org/10.52436/1.jutif.2024.5.4.1492

Arunkumar, K. A., Kriti, K., Das, A., & Bhattacharyya, B. (2019). Wireless Speakers using WiFi and IoT. *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*, 1–6.

Bora, J., Dehingia, S., Boruah, A., Chetia, A. A., & Gogoi, D. (2022). *Real-time* Assamese Sign Language Recognition using MediaPipe and Deep Learning. *Procedia Computer Science*, *218*(2022), 1384–1393. https://doi.org/10.1016/j.procs.2023.01.117

Dewi, S., Ramadhani, F., & Djasmayena, S. (2024). Klasifikasi Jenis Jerawat Berdasarkan Gambar Menggunakan Algoritma CNN (Convolutional Neural Network). *Hello World Jurnal Ilmu Komputer*, *3*(2), 68–73. https://doi.org/10.56211/helloworld.v3i2.518

Godase, V., Modi, S., Misal, V., & Kulkarni, S. (2025). LoRaEdge-ESP32 Synergy: Revolutionizing Farm Weather Data Collection with Low-Power, Long-Range IoT. *Advance Research in Analog and Digital Communications*, *2*(2), 1–11.

Guo, Z., Peng, J., Luo, B., Ma, A., Zheng, X., & Chip, L. M. (2024). Design , Analysis , and Implementation of a Frequency Modulation Receiver System with Enhanced Audio Power Amplification. *International Core Journal of Engineering*, *10*(5), 463–470. https://doi.org/10.6919/ICJE.202405

I Gusti Agung Made Yoga Mahaputra, Putri Alit Widyastuti Santiary, & I Ketut Swardika. (2025). Rancang Bangun Penerjemah BISINDO *Real-time* Berbasis Kamera dan Deep Learning dengan Kendali Suara ESP32 WiFi. *Jurnal Elektro Dan Mesin Terapan*, *11*(1), 33–42. https://doi.org/10.35143/elementer.v11i1.6578

Kumar, M., Yadav, D. K., Ray, S., & Tanwar, R. (2024). Handling illumination variation for motion detection in video through intelligent method: An application for smart surveillance system. *Multimedia Tools and Applications*, *83*(10), 29139–29157.

Nisria, Mustafa, & Hadis. (2022). IMPLEMENTASI BISINDO DALAM BERKOMUNIKASI PADA SESAMA ANAK TUNARUNGU (Implementation of BISINDO in communicating with deaf children). *Pinisi Journal of Education*, 1–10.

Putra, I. A., Nurhayati, O. D., & Eridani, D. (2022). Human Action Recognition (HAR) Classification Using MediaPipe and *Long Short-Term Memory* (LSTM). *Teknik*, *43*(2), 190–201. https://doi.org/10.14710/teknik.v43i2.46439

Sánchez-Brizuela, G., Cisnal, A., de la Fuente-López, E., Fraile, J. C., & Pérez-Turiel, J. (2023). Lightweight *real-time* hand segmentation leveraging MediaPipe *landmark* detection. *Virtual Reality*, *27*(4), 3125–3132. https://doi.org/10.1007/s10055-023-00858-0

Sari, I., Fivrenodi, Altiarika, E., & Sarwindah. (2023). Sistem Pengembangan Bahasa Isyarat Untuk Berkomunikasi dengan Penyandang Disabilitas (Tunarungu). *Journal of Information Technology and Society*, *1*(1), 20–25. https://doi.org/10.35438/jits.v1i1.21

Sebastian, C., Limanza, J., Laurentia, L., Harefa, J., Sebastian, C., Limanza, J., Laurentia, L., & Harefa, J. (2025). Indonesian Sign Language (BISINDO) Recognition Using Indonesian Sign Language (BISINDO) Recognition Using Spatially Aware Body *Gesture* Recognition Spatially Aware Body *Gesture* Recognition. *Procedia Computer Science*, *269*, 1002–1011. https://doi.org/10.1016/j.procs.2025.09.042

Ur Rehman, M., Ahmed, F., Khan, M. A., Tariq, U., Alfouzan, F. A., Alzahrani, N. M., & Ahmad, J. (2022). Dynamic hand *gesture* recognition using 3D-CNN and LSTM networks. *Computers, Materials and Continua*, *70*(3), 4675–4690. https://doi.org/10.32604/cmc.2022.019586

Wang, C., & Yan, J. (2023). A Comprehensive Survey of RGB-Based and Skeleton-Based Human Action Recognition. *IEEE Access*, *11*, 53880–53898. https://doi.org/10.1109/ACCESS.2023.3282311