e-ISSN: 2581-0545 - https://journal.itera.ac.id/index.php/jsat/

Perbandingan Metode CNN-LSTM dan CNN-GRU dalam Memprediksi Gesture Tangan

Received 00th January 20xx
Accepted 00th Febuary 20xx
Published 00th March 20xx

Open Access

DOI: 10.35472/x0xx0000

Veni Zahara Kartika[*a], Yosia Adwily Nainggolan[b], M. Rizki Alfaina[c], Tri Murniya Ningsih[d], Ima Alifah Izati Zalfa[e], Anita Rahma Pramoda Cahyani[f], Ade Lailani[g]

[a, b, d, e, f, g] *Program Studi Sains Data, Fakultas Sains, Institut Teknologi Sumatera*

[c] *Program Studi Teknik Informatika, Fakultas Teknologi Industri, Institut Teknologi Sumatera*

** Corresponding E-mail: veni.121450075@student.itera.ac.id*

Abstract: Hand gesture recognition is one of the important areas in non-verbal communication and computer vision technology that continues to develop. This research aims to compare the performance of two deep learning methods, namely Convolutional Neural Network with Long Short-Term Memory (CNN-LSTM) and Convolutional Neural Network with Gated Recurrent Unit (CNN-GRU) in identifying hand movements. The dataset used is the Jester Dataset which consists of 148,092 videos with 27 labels, which are then simplified into 5 main labels. The preprocessing process includes data subsampling, frame resizing to 64x64 pixels, conversion to grayscale, and normalization. Both models were trained using the Adam optimizer with the Sparse Categorical Cross Entropy loss function for 10 epochs. The research results show that the CNN-GRU model is slightly superior with an accuracy of 76% in validation and 70% in testing, compared to CNN-LSTM which achieved 82% in validation and 60% in testing. While both models performed well on specific gestures such as 'Finger Tapping' and 'Two Fingers Slide Down', they still experienced challenges in classifying the 'Two Fingers Move Away' gesture. This research contributes to the development of hand gesture recognition technology and provides new insights into gesture-based human-computer interaction.

Keywords: *Deep Learning, Convolutional Neural Network, Gated Recurrent Unit, Long Short-Term Memory*

Abstrak: Pengenalan gestur tangan merupakan salah satu bidang penting dalam komunikasi non-verbal dan teknologi computer vision yang terus berkembang. Penelitian ini bertujuan membandingkan kinerja dua metode deep learning, yaitu *Convolutional Neural Network dengan Long Short-Term Memory* (CNN-LSTM) dan *Convolutional Neural Network* dengan *Gated Recurrent Unit* (CNN-GRU) dalam mengidentifikasi gerakan tangan. Dataset yang digunakan adalah Jester Dataset yang terdiri dari 148.092 video dengan 27 label, yang kemudian disederhanakan menjadi 5 label utama. Proses preprocessing meliputi subsampling data, resize frame menjadi 64x64 piksel, konversi ke grayscale, dan normalisasi. Kedua model dilatih menggunakan optimizer Adam dengan fungsi *loss Sparse Categorical Cross Entropy* selama 10 epoch. Hasil penelitian menunjukkan bahwa model CNN-GRU sedikit lebih unggul dengan akurasi 76% pada validasi dan 70% pada pengujian, dibandingkan CNN-LSTM yang mencapai 82% pada validasi dan 60% pada pengujian. Meskipun kedua model memiliki performa baik pada gerakan spesifik seperti 'Mengetuk Jari' dan 'Menggeser Dua Jari Ke Bawah', keduanya masih mengalami tantangan dalam mengklasifikasikan gerakan 'Menjauhkan Dua Jari'. Penelitian ini memberikan kontribusi dalam pengembangan teknologi pengenalan gestur tangan dan memberikan wawasan baru dalam interaksi manusia-komputer berbasis gerakan.

Kata Kunci : *Deep Learning, Convolutional Neural Network, Gated Recurrent Unit, Long Short Term Memory.*

e-ISSN: 2581-0545 - https://journal.itera.ac.id/index.php/jsat/

Introduction / Pendahuluan

Latar Belakang

Komunikasi adalah elemen fundamental dalam kehidupan manusia yang memungkinkan individu untuk berinteraksi, berbagi informasi, dan membangun hubungan sosial. Dalam konteks sosial, kemampuan untuk berkomunikasi secara efektif sangat penting untuk keberhasilan dalam berbagai aspek kehidupan, termasuk pendidikan, pekerjaan, dan interaksi sehari-hari [1]. Komunikasi biasanya melalui cara verbal dan non-verbal yang melibatkan penyampaian pesan antar pihak yang dapat diterima oleh pihak lain [2]. Pengenalan gestur tangan termasuk kedalam komunikasi non-verbal yang digunakan secara alami untuk berkomunikasi dalam lingkungan sekitar [3].

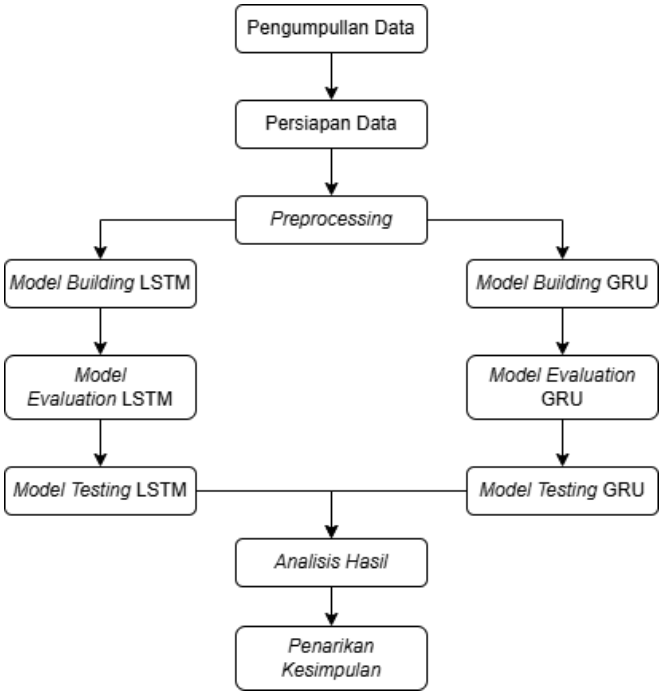
Dengan kemajuan teknologi, khususnya dalam bidang *machine learning* dan *computer vision*, banyak inovasi telah diperkenalkan untuk meningkatkan kehidupan manusia. Teknologi ini tidak hanya mencakup pengenalan wajah dan sistem keamanan berbasis visual, tetapi juga mencakup sistem yang dirancang untuk mengenali gestur tangan. Model *deep learning*, seperti *Convolutional Neural Network* (CNN), telah terbukti efektif dalam mengekstraksi fitur dari data visual, sementara arsitektur *Long Short-Term Memory* (LSTM) dan *Gated Recurrent Unit* (GRU) sangat baik dalam menangkap hubungan temporal dalam data berurutan.

Kombinasi CNN-LSTM dan CNN-GRU menjadi pendekatan yang menjanjikan dalam pengenalan gestur tangan seperti penelitian yang dilakukan oleh G. B. Prananta dkk yang berjudul "Deteksi Pengenalan Gestur Tangan Secara Real-Time Menggunakan Jaringan Syaraf Tiruan Konvolusional" menggunakan model CNN dengan total dataset yang digunakan adalah 634 gestur tangan memberikan hasil bahwa sistem dapat mendeteksi gerakan tangan. Lalu pada penelitian yang dilakukan oleh R. Bi dengan judul *"Sensor-based Gesture Recognition With Convolutional Neural Network",* menggunakan CNN menghasilkan aproksimasi 97% dan akurasi sebesar 72%. Temuan dari penelitian terdahulu ini menunjukkan bahwa CNN memberikan hasil yang optimal dalam pengenalan gerak tangan. Namun pada penelitian ini terdapat pengembangan model yang menggabungkan CNN dengan *Long Short-Term Memory* (LSTM) dan *Gated Recurrent Unit* (GRU) dengan harapan model dapat menghasilkan peningkatan akurasi pengenalan gestur tangan secara signifikan.

Dengan demikian penelitian ini berfokus pada perbandingan akurasi model CNN-LSTM dan CNN-GRU dalam mengidentifikasi gestur tangan berdasarkan data video serta menentukan model mana yang lebih unggul diantara kedua model yang digunakan. Harapannya penelitian ini dapat memberikan kontribusi dalam pengembangan teknologi pengenalan gestur tangan serta memberikan ide baru yang lebih luas dalam interaksi manusia dengan komputer di masa depan.

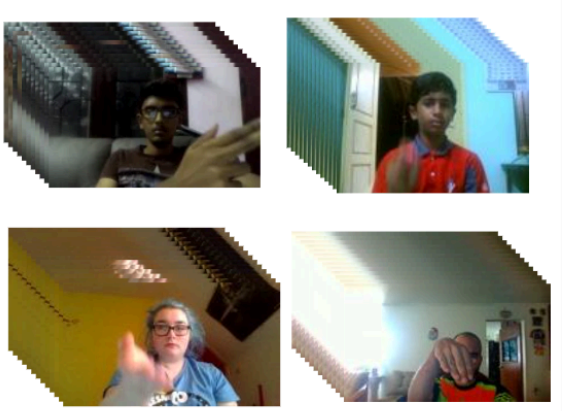
Method / Metode

Pada penelitian ini digunakan metode gabungan CNN dan LSTM dengan gabungan CNN dan GRU dalam memprediksi gerakan tangan. Tahapan yang dilakukan pada penelitian ini terbagi dalam 4 tahapan yaitu, melakukan preprocessing pada dataset, membuat masing-masing model pada tahapan model building, dan mengevaluasi model yang telah dibuat. Setelah keempat tahapan tersebut telah selesai, maka didapatkan nilai akurasi, recall, dan precision dalam proses validasi. Selanjutnya, proses testing dilakukan untuk mengetahui ketepatan prediksi label berdasarkan model yang telah dikembangkan. Nilai validasi tersebut dianalisis oleh penulis untuk mengetahui metode mana yang lebih baik. Ketepatan Prediksi juga digunakan sebagai pertimbangan untuk mengukur metode mana yang lebih baik. Alur penelitian yang dilakukan oleh penulis seperti pada **Gambar 1.**



Gambar 1. *Diagram Alir Kerangka Penelitian.*

Persiapan Data



Gambar 2. *Dataset*

Data yang digunakan pada penelitian ini adalah video seperti pada **Gambar 2.** yang bersumber dari penelitian yang dilakukan oleh Joanna Materzynska dkk dengan judul *The Jester Dataset: A Large-Scale Video Dataset of Human Gestures* yang datasetnya dapat diunduh melalui laman Kaggle [5]. Video yang digunakan terdiri dari 148.092 video dengan durasi 3 detik, dengan total 5.331.312 frame pada keseluruhan video. Pada setiap video yang direkam, terdapat 27 label dengan informasi detail tentang spesifikasi dataset dapat dilihat pada **Tabel 1**.

Table 1. *Distribusi Label dataset*

Jumlah Label	Label
4374	Doing other things
1859	Pulling Two Fingers In
1847	Zooming Out With Two Fingers
1844	No gesture
1843	Pushing Two Fingers Away
1841	Thumb Up
1832	Sliding Two Fingers Down
1832	Zooming Out With Full Hand
1829	Pulling Hand In
1824	Swiping Down
1821	Stop Sign
1818	Drumming Fingers
1816	Sliding Two Fingers Left
1812	Pushing Hand Away
1810	Thumb Down
1801	Zooming in With Two Fingers
1799	Zooming In With Full Hand
1789	Shaking Hand
1788	Rolling Hand Forward
1780	Sliding Two Fingers Right
1779	Sliding Two Fingers Up
1768	Swiping Up
1762	Swiping Left
1730	Swiping Right
1715	Rolling Hand Backward
1380	Turning Hand Counterclockwise
1327	Turning Hand Clockwise

Preprocessing

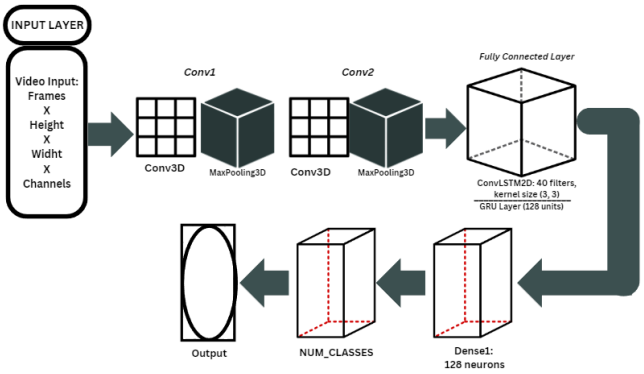
Gambar 3. *Diagram Alir Preprocessing*

Pada **Gambar 3**, merupakan proses preprocessing yang dimulai dari data subsampling, yaitu proses resampling data yang semula 27 label menjadi 5 label dan setiap labelnya memiliki 1000 folder yang berisi 30 frame pada setiap videonya. Label yang digunakan adalah 1000 berlabel 'Menjauhkan Tangan' yang diterjemahkan dari 'Pushing Hand Away', 1000 video berlabel 'Mengetuk Jari' yang diterjemahkan dari 'Drumming Fingers', 1000 video berlabel 'Melakukan hal lain' yang diterjemahkan dari 'Doing other things', 1000 video berlabel 'Menjauhkan Dua Jari' yang diterjemahkan dari 'Pushing Two Fingers Away', dan 1000 video berlabel 'Menggeser Dua Jari Ke Bawah' yang diterjemahkan dari 'Sliding Two Fingers Down'. Setelah proses subsampling, video tersebut akan dilakukan frame resize yang sebelumnya memiliki dimensi foto yang bervariasi. Foto tersebut diubah dimensinya menjadi 64x64 agar performa model yang akan dibuat meningkat.

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Selanjutnya, adalah mengubah gambar yang telah diubah dimensinya menjadi grayscale. Langkah terakhir preprocessing pada penelitian ini adalah normalisasi dimensi foto yang akan diolah. Dimensi foto 64x64x3 diubah menjadi 64x64x1 agar kinerja model dapat meningkat.

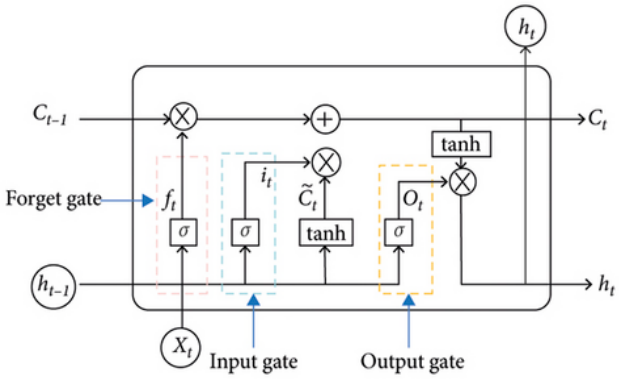
Arsitektur Model



Gambar 4. *Arsitektur Model*

a. CNN-LSTM

Arsitektur yang digunakan pada penelitian ini adalah CNN-LSTM untuk prediksi gerakan tangan pada data train. CNN sebagai lapisan konvolusional digunakan untuk ekstraksi fitur pada pengolahan data masukan. *Long Short Term Memory* (LSTM) adalah pengembangan *neural network* yang dapat digunakan untuk pemodelan data *time series*. LSTM digunakan untuk ekstraksi informasi dimensi gambar yang telah dilakukan proses konvolusional. Sebuah Penelitian "Klasifikasi Data Penderita Skizofrenia Menggunakan CNN-LSTM dan CNN-GRU pada Data Sinyal EEG 2D" mengatakan bahwa penggabungan CNN-LSTM berhasil diterapkan dalam penggunaan data sekuensial. Cell LSTM adalah bagian penting dari jaringan saraf LSTM yang berfungsi seperti sebuah memori. Memori ini sangat fleksibel, bisa menyimpan informasi dalam waktu yang lama atau singkat, dan bisa memilih informasi yang paling relevan untuk menghasilkan output yang sesuai. Kemampuan inilah yang membuat LSTM unggul dalam memproses data yang memiliki urutan atau ketergantungan waktu. Berikut merupakan arsitektur LSTM pada **Gambar 5 .** beserta persamaan matematis yang terdiri dari *forget gate*, *input gate*, dan *output gate*.



Gambar 5. *Arsitektur LSTM*

$$f_t = \sigma(W_f \cdot [h_{t-1}, \chi_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, \chi_t] + b_i) \quad (2)$$

$$\overline{C}_t = tanh(W_c \cdot [h_{t-1}, \chi_t] + b_c) \quad (3)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \overline{C} \quad (4)$$

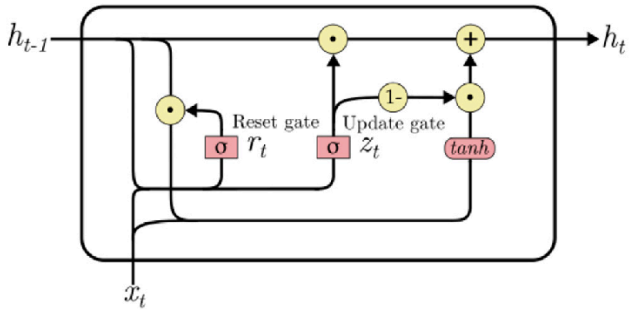
$$o_t = \sigma(W_o \cdot [h_{t-1}, \chi_t] + b_o) \quad (5)$$

$$h_t = o_t \cdot tanh(C_t) \quad (6)$$

b. CNN-GRU

Penggunaan gabungan *Convolutional Neural Network* (CNN) dan *Gated Recurrent Units* (GRU), yang dikenal sebagai CNN-GRU, merupakan pendekatan yang sangat efektif dalam pengenalan *gesture* tangan. Dalam model ini, CNN bertugas untuk mengekstraksi fitur spasial dari gambar tangan atau gerakan tangan. CNN dapat mengenali elemen-elemen dasar dari gambar, seperti bentuk jari, orientasi tangan, dan posisi relatif tangan dalam ruang. Teknik ini sangat berguna dalam menangani gambar yang memiliki kompleksitas visual, seperti gestur tangan. [7]

Setelah CNN mengidentifikasi fitur-fitur penting dalam gambar, GRU digunakan untuk menangani aspek temporal atau urutan waktu. GRU, yang merupakan jenis *Recurrent Neural Network* (RNN), lebih efisien dalam memproses urutan data dibandingkan dengan LSTM, karena GRU lebih ringan dan cepat dalam pelatihan. GRU sangat efektif dalam menangkap dinamika gerakan tangan yang terjadi seiring waktu, yang sangat penting dalam pengenalan *gesture* tangan yang bergantung pada urutan gerakan.[8] Berikut arsitektur **Gambar 6.** dan persamaan matematis dalam model GRU yang terdiri dari *reset gate* dan *update gate*.



Gambar 6. *Arsitektur GRU*

$$R_t = \sigma(X_t W_{xr} + H_{t-1} W_{hr} + b_r) \quad (7)$$

$$Z_t = \sigma(X_t W_{xz} + H_{t-1} W_{hz} + b_z) \quad (8)$$

Gabungan CNN dan GRU ini memungkinkan model untuk tidak hanya mengenali pola visual dari gambar tangan, tetapi juga memahami bagaimana gerakan tangan berubah seiring waktu. Ini memberikan keuntungan besar dalam aplikasi pengenalan gesture, di mana analisis spasial dan temporal harus dilakukan secara bersamaan. Model CNN-GRU memberikan hasil yang lebih baik dalam mengidentifikasi *gesture* tangan, bahkan ketika gerakan dilakukan dalam urutan yang kompleks atau dengan variasi kecepatan dan arah.

Proses Pelatihan

Kedua model dilatih menggunakan *optimizer* Adam dengan fungsi *loss Space Categorical Cross Entropy*. Pelatihan dilakukan selama 10 epoch dengan batch size sebesar 32. Dataset dibagi menjadi data pelatihan dan validasi untuk memantau performa model selama proses pelatihan.

Evaluasi

Evaluasi model dilakukan menggunakan beberapa metrik:

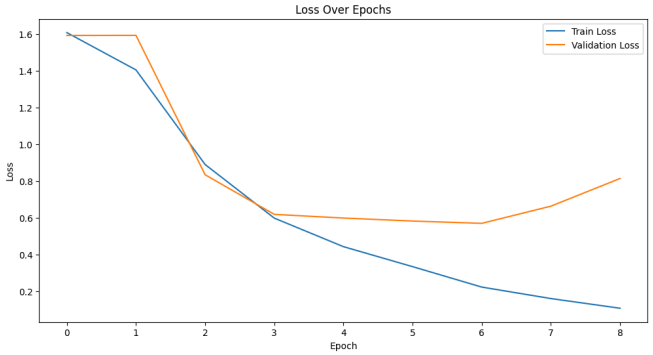
- Akurasi : Mengukur persentase prediksi yang benar.
- Precision, Recall, F1-Score: Digunakan untuk mengevaluasi performa model di setiap kelas secara mendalam.
- Confusion Matrix: Menunjukkan distribusi prediksi model terhadap label sebenarnya.

Hasil evaluasi dari kedua model dibandingkan untuk menentukan arsitektur mana yang lebih unggul dalam mengenali gestur tangan. Berdasarkan penelitian sebelumnya, CNN-LSTM diharapkan memberikan hasil lebih baik karena kemampuannya menangkap pola temporal jangka panjang dengan lebih baik dibandingkan GRU yang lebih sederhana namun efisien secara komputasi.

Hasil dan Pembahasan

CNN-LSTM

Pada penelitian ini, dilakukan percobaan terhadap data *hand gesture recognition* dengan data video *hand gesture* yang di resampling menjadi 5 label dan setiap labelnya memiliki 1000 folder berisi 30 frame dari sebuah video. Data ini kemudian diolah menggunakan model *combined* yakni CNN-LSTM dan CNN-GRU. Model *Combined* CNN-LSTM memiliki arsitektur model yang dirancang dengan beberapa lapisan, yaitu 2 lapisan konvolusi dengan jumlah filter 32 pada lapisan pertama dan 64 filter pada lapisan kedua, dengan menggunakan fungsi aktivasi *ReLU* untuk mengekstrak fitur penting dari *frame* video dan meningkatkan non-linieritas pada model, 2 lapisan *pooling* untuk mereduksi dimensi data dan kompleksitas komputasi tanpa kehilangan informasi pentingnya, selanjutnya lapisan LSTM konvolusional untuk menangkap dependensi temporal dalam urutan video. *Output* dari lapisan LSTM konvolusional ini diratakan menggunakan lapisan *flatten* menjadi vektor tunggal kemudian diteruskan ke lapisan *fully connected* untuk mengintegrasikan informasi dari semua fitur dengan 128 unit dan fungsi aktivasi *ReLU,* dan akhirnya menghasilkan lapisan *output* dengan fungsi aktivasi *softmax* untuk memberikan probabilitas pada setiap kelas. Model ini dikompilasi dengan menggunakan *optimizer Adam* dan fungsi *loss* dengan menggunakan *Sparse Categorical Cross Entropy.* Kemudian, dilakukan pelatihan model dengan data *train* dan data validasi, proses latih dilakukan selama 10 epoch dengan *batch size* sebesar 32 dan tambahan *early stopping*.

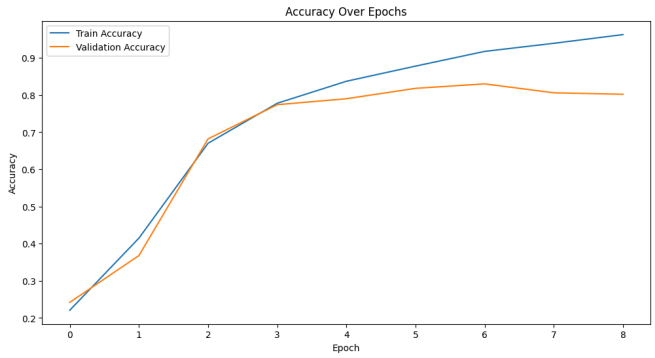


Gambar 7. *Plot Loss terhadap Epochs pada Model CNN-LSTM.*

Gambar 7. menunjukkan plot *loss* terhadap *epoch* untuk data *train* cenderung mengalami penurunan secara keseluruhan sedangkan untuk data validasi mengalami

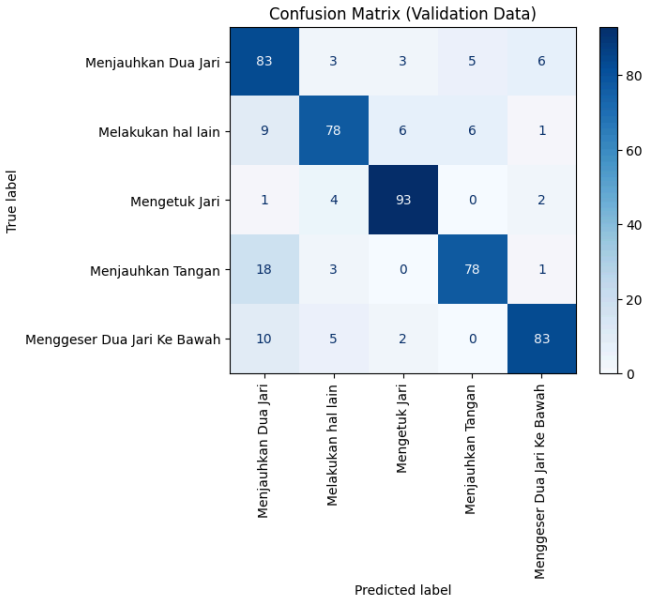
Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

penurunan namun pada *epoch* ke-6 loss mengalami kenaikan dan belum signifikan. Hasil pelatihan model menunjukkan bahwa seiring bertambahnya *epoch, train loss* menurun dari 1.6176 hingga 0.1152 dan *validation loss* cenderung menurun dari 1.5931 hingga 0.8151, ini menunjukkan model semakin baik dalam memprediksi data. Pelatihan model CNN-LSTM berhenti hingga epoch ke-8 dikarenakan tidak adanya peningkatan validation loss dalam 2 epoch.



Gambar 8. *Plot Akurasi terhadap Epochs pada Model CNN-LSTM.*

Berdasarkan **Gambar 8.** menunjukkan plot akurasi terhadap *epoch* cenderung meningkat secara keseluruhan. Setiap *epoch, train accuracy* meningkat dari 0.1985 hingga 0.9600 dan *validation accuracy* cenderung meningkat dan stabil pada beberapa *epoch* dari 0.2420 hingga 0.8020, ini menunjukkan model semakin baik dalam mengenali pola dan informasi dari data.



Gambar 9. *Confusion Matrix Model CNN-LSTM (Validation Data).*

Berdasarkan **Gambar 9.** *confusion matrix* menunjukkan hasil prediksi kinerja model terhadap data validasi dengan data asli yang memiliki 500 *video-clips*. Ini memprediksi model terhadap 5 kelas label. Pada diagonal utamanya, model memprediksi 83 video berlabel 'Menjauhkan Dua Jari', 78 video berlabel 'Melakukan hal lain', 93 video berlabel 'Mengetuk Jari', 78 video berlabel 'Menjauhkan Tangan', dan 83 video berlabel 'Menggeser Dua Jari Ke Bawah' yang berhasil memprediksi dengan benar dan sisanya di luar diagonal menunjukkan kesalahan prediksi yang diklasifikasikan ke label lain.

Table 2. *Evaluasi Model CNN-LSTM pada Data Validation.*

Label	*Precision*	*Recall*	*F1-Score*	*Support*
Menjauhkan Dua Jari	*0.69*	*0.83*	*0.75*	*100*
Melakukan hal lain	*0.84*	*0.78*	*0.81*	*100*
Mengetuk Jari	*0.89*	*0.93*	*0.91*	*100*
Menjauhkan Tangan	*0.88*	*0.78*	*0.83*	*100*
Menggeser Dua Jari Ke Bawah	*0.89*	*0.83*	*0.86*	*100*
accuracy			*0.83*	*500*
macro avg	*0.84*	*0.83*	*0.83*	*500*
weighted avg	*0.84*	*0.83*	*0.83*	*500*

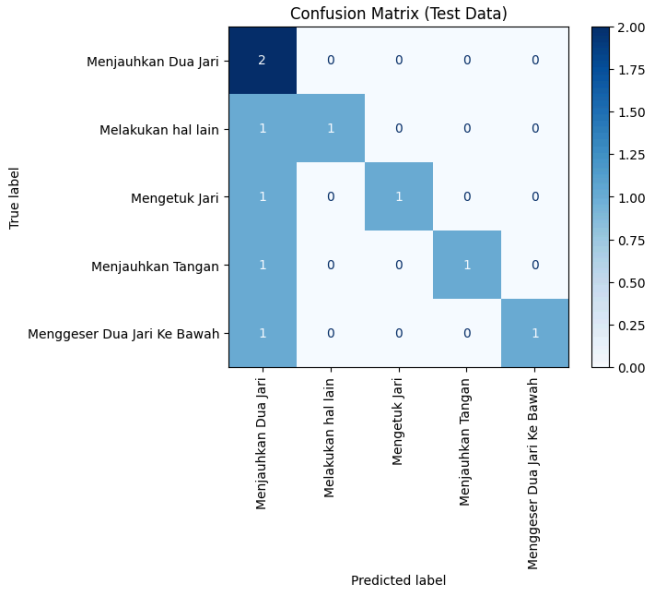
Berdasarkan **Table 2.** Hasil evaluasi model menunjukkan model mencapai akurasi keseluruhan sebesar 83% yang menunjukkan performa yang cukup baik dalam klasifikasi. Dengan nilai rata-rata kisaran 83% untuk *precision, recall,* dan *f1-score* pada *macro average* maupun *weighted average.* Pada label 'Mengetuk Jari' dan 'Menggeser Dua Jari Ke Bawah' menunjukkan keseimbangan performa terbaik, yang masing-masing nilai *f1-score*-nya 0.91 dan 0.83, yang menunjukkan model baik dalam mengenali dan memprediksi label. Sebaliknya pada label 'Menjauhkan Dua Jari' memiliki nilai *f1-score* terendah yaitu 0.75, menunjukkan ketidakseimbangan performa model dan adanya ruang untuk perbaikan dalam membedakan label.

Table 3. *Hasil Prediksi Label Data Test Model CNN-LSTM .*

video_id	predicted_label
109	Menjauhkan Dua Jari
20	Menjauhkan Dua Jari

233	Mengetuk Jari
79	Menjauhkan Dua Jari
85	Menggeser Dua Jari Ke Bawah
3	Menjauhkan Dua Jari
1	Melakukan hal lain
14	Menjauhkan Tangan
6	Menjauhkan Dua Jari
11	Menjauhkan Dua Jari

Berdasarkan **Table 3.** ini memberikan hasil prediksi model untuk kategori gerakan yang dilakukan dari setiap video dalam data test yang terdiri dari 10 video. Label prediksi menunjukkan hasil prediksi model terhadap data yang belum pernah ditemui sebelumnya tentang gerakan yang dilakukan. Hasil ini dapat digunakan untuk mengevaluasi kinerja model yang dilatih.



Gambar 10. *Confusion Matrix Model CNN-LSTM (Test Data).*

Gambar 10. menunjukkan *confusion matrix* hasil prediksi kinerja model terhadap data test dengan data asli yang memiliki 10 *video-clips*. Ini memprediksi model terhadap 5 kelas label. Pada diagonal utamanya, model memprediksi 2 video berlabel 'Menjauhkan Dua Jari', 1 video berlabel 'Melakukan hal lain', 1 video berlabel 'Mengetuk Jari', 1 video berlabel 'Menjauhkan Tangan', dan 1 video berlabel 'Menggeser Dua Jari Ke Bawah' yang berhasil memprediksi dengan benar dan sisanya di luar diagonal menunjukkan kesalahan prediksi yang diklasifikasikan ke label lain.

Table 4. *Evaluasi Model CNN-LSTM pada Data Validation.*

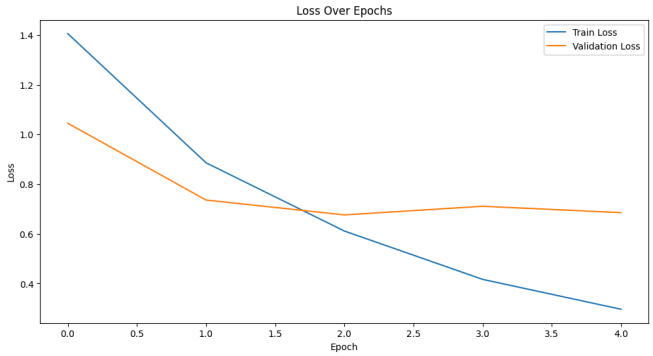
Label	*Precision*	*Recall*	*F1-Score*	*Support*
Menjauhkan Dua Jari	*0.33*	*1.00*	*0.50*	*2*
Melakukan hal lain	*1.00*	*0.50*	*0.67*	*2*
Mengetuk Jari	*1.00*	*0.50*	*0.67*	*2*
Menjauhkan Tangan	*1.00*	*0.50*	*0.67*	*2*
Menggeser Dua Jari Ke Bawah	*1.00*	*0.50*	*0.67*	*2*
accuracy			*0.60*	*10*
macro avg	*0.50*	*0.60*	*0.63*	*10*
weighted avg	*0.50*	*0.60*	*0.63*	*10*

Berdasarkan **Tabel 4.** akurasi model dalam memprediksi data test yang belum pernah ditemui adalah 60%. Hal ini menunjukkan model cukup baik dalam memberi label pada data yang belum pernah ditemui dan mengeneralisasinya sesuai dengan label aktual.

CNN-GRU

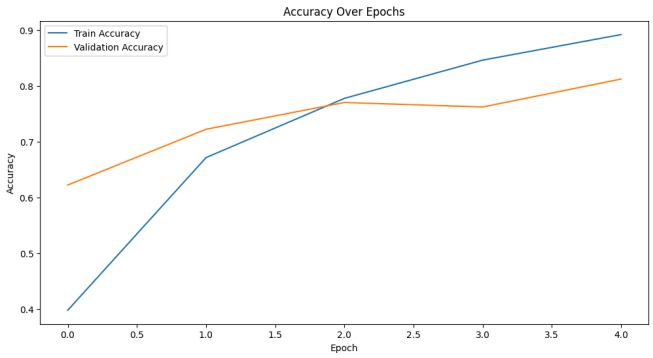
Pada model CNN-GRU, arsitektur model yang digunakan terdiri dari dua lapisan konvolusi yakni lapisan pertama menggunakan 32 filter dan lapisan kedua menggunakan 64 filter disertai dengan fungsi aktivasi ReLU untuk mengekstrak fitur penting dari frame video dan membuat model dapat mempelajari pola non-linear, lapisan *pooling* untuk mereduksi dimensi data dengan mengambil nilai maksimum dalam pixel. Selanjutnya, model dihubungkan ke lapisan GRU untuk menangkap hubungan temporal dalam fitur yang diekstraksi CNN sebelumnya berupa gambar berurut dari kumpulan video. Output dari lapisan GRU dihubungkan menggunakan lapisan *flatten* kemudian diteruskan ke lapisan *fully connected* dan menghasilkan prediksi akhir. Model ini dikompilasi dengan menggunakan optimizer *Adam* dan fungsi *loss* dengan menggunakan *Sparse Categorical Cross Entropy.* Setelah model terdefinisi maka dilakukan pelatihan model dengan data train dan data validasi, proses pelatihan dilakukan dengan epoch sebanyak 10 dan *batch size* sebesar 32 dan tambahan *early stopping*.

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia



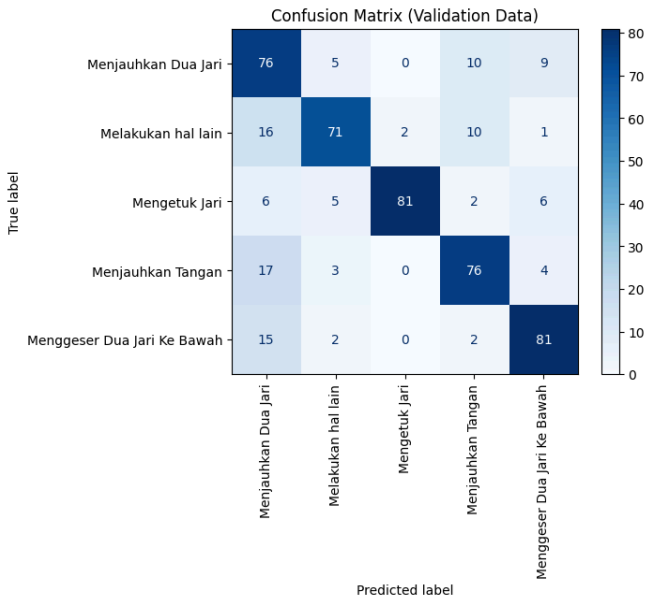
Gambar 11. *Plot Loss terhadap Epochs pada Model CNN-GRU.*

Berdasarkan **Gambar 11.** plot loss terhadap epoch untuk data train cenderung mengalami penurunan secara konsisten sedangkan untuk data validasi mengalami penurunan di awal namun pada epoch ke-3 loss mengalami kenaikan. Hasil pelatihan model menunjukkan bahwa seiring bertambahnya *epoch, train loss* menurun dari 1.5642 hingga 0.3120 dan *validation loss* cenderung menurun dari 1.0450 hingga 0.6852. Hal ini menunjukkan model semakin baik dalam memprediksi dan mempelajari data latih, sedangkan validation loss menunjukkan bahwa model menggeneralisasi data validasi dengan baik. Pelatihan model CNN-GRU berhenti hingga epoch ke-4 dikarenakan tidak adanya peningkatan validation loss dalam 2 epoch.



Gambar 12. *Plot Akurasi terhadap Epochs pada Model CNN-GRU.*

Gambar 12. menunjukkan plot cenderung meningkat secara keseluruhan. Setiap *epoch, train accuracy* meningkat dari 0.2895 hingga 0.8881 dan *validation accuracy* cenderung meningkat dari 0.6220 hingga 0.8120, menunjukkan model semakin baik dalam mengenali pola dan informasi dari data.



Gambar 14. *Confusion Matrix Model CNN-GRU (Validation Data).*

Gambar 13. menunjukkan hasil prediksi model terhadap data validasi dengan data asli yang memiliki 500 video-clips. Model dapat memprediksi 76 video berlabel 'Menjauhkan Dua Jari', 71 video berlabel 'Melakukan hal lain', 81 video berlabel 'Mengetuk Jari', 76 video berlabel 'Menjauhkan Tangan', dan 81 video berlabel 'Menggeser Dua Jari Ke Bawah' yang berhasil memprediksi dengan benar dan sisanya di luar diagonal menunjukkan kesalahan prediksi yang diklasifikasikan ke label lain.

Table 5. *Evaluasi Model CNN-GRU pada Data Validation.*

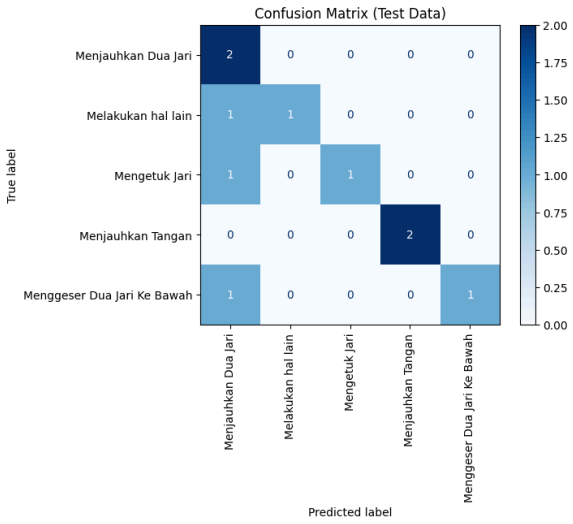
Label	***Precision***	***Recall***	***F1-Score***	***Support***
Menjauhkan Dua Jari	*0.58*	*0.76*	*0.66*	*100*
Melakukan hal lain	*0.83*	*0.71*	*0.76*	*100*
Mengetuk Jari	*0.98*	*0.81*	*0.89*	*100*
Menjauhkan Tangan	*0.76*	*0.76*	*0.76*	*100*
Menggeser Dua Jari Ke Bawah	*0.80*	*0.81*	*0.81*	*100*
accuracy			*0.77*	*500*
macro avg	*0.79*	*0.77*	*0.78*	*500*
weighted avg	*0.79*	*0.77*	*0.78*	*500*

Berdasarkan **Table 5.** model memiliki akurasi sebesar 77%, dengan rata-rata precision, recall, dan F1-score yang juga berada di kisaran 77% untuk rata-rata macro maupun weighted. Kelas 'Mengetuk Jari' dan 'Menggeser Dua Jari Ke Bawah' menunjukkan performa terbaik dengan F1-score masing-masing 0.89 dan 0.81, menunjukkan bahwa model sangat baik dalam mengidentifikasi instance dari kelas tersebut. Sebaliknya, kelas 'Menjauhkan Dua Jari' memiliki performa terburuk dengan F1-score sebesar 0.66, menunjukkan bahwa model kesulitan mengenali sebagian besar instance kelas tersebut.

Table 6. *Hasil Prediksi Label Data Test Model CNN-GRU.*

video_id	predicted_label
109	Menjauhkan Dua Jari
20	Menjauhkan Dua Jari
233	Mengetuk Jari
79	Menjauhkan Tangan
85	Menjauhkan Dua Jari
3	Menjauhkan Dua Jari
1	Melakukan hal lain
14	Menjauhkan Tangan
6	Menjauhkan Dua Jari
11	Menggeser Dua Jari Ke Bawah

Tabel 6. merupakan hasil prediksi label menggunakan *data test* sebanyak 10 video. Label prediksi menunjukkan hasil prediksi model terhadap data yang belum pernah ditemui sebelumnya tentang gerakan yang dilakukan.



Gambar 14. *Confusion Matrix Model CNN-GRU (Test Data).*

Gambar 14. menunjukkan distribusi hasil prediksi label terhadap data test dan dapat diketahui bahwa model memprediksi 2 video berlabel 'Menjauhkan Dua Jari', 1 video berlabel 'Melakukan hal lain', 1 video berlabel 'Mengetuk Jari', 2 video berlabel 'Menjauhkan Tangan', dan 1 video berlabel 'Menggeser Dua Jari Ke Bawah' yang berhasil memprediksi dengan benar dan sisanya di luar diagonal menunjukkan kesalahan prediksi yang diklasifikasikan ke label lain.

Table 7. *Evaluasi Model CNN-GRU pada Data Validation.*

Label	*Precision*	*Recall*	*F1-Score*	*Support*
Menjauhkan Dua Jari	*0.40*	*1.00*	*0.57*	*2*
Melakukan hal lain	*1.00*	*0.50*	*0.67*	*2*
Mengetuk Jari	*1.00*	*0.50*	*0.67*	*2*
Menjauhkan Tangan	*1.00*	*1.00*	*1.00*	*2*
Menggeser Dua Jari Ke Bawah	*1.00*	*0.50*	*0.67*	*2*
accuracy			*0.70*	*10*
macro avg	*0.88*	*0.70*	*0.71*	*10*
weighted avg	*0.88*	*0.70*	*0.71*	*10*

Berdasarkan **Tabel 7.** akurasi model dalam memprediksi data test yang belum pernah ditemui adalah 70%. Hal ini menunjukkan model cukup baik dalam memberi label pada data yang belum pernah ditemui, namun pada proses pelatihan model menunjukkan indikasi overfitting dan ketika dilakukan prediksi menggunakan data test, model tetap dapat memprediksi gerakan tangan dengan baik dan menghasilkan akurasi yang lebih tinggi.

Table 8. *Evaluasi Akurasi Model CNN-LSTM dan CNN-GRU*

Evaluasi	*CNN-LSTM*	*CNN-GRU*
Data Train	*0.94*	*0.83*
Data Validation	*0.82*	*0.76*
Data Test	*0.60*	*0.70*

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Hasil perbandingan antara kedua model berdasarkan **Tabel 8.** menunjukkan bahwa model CNN-GRU memiliki keunggulan pada *data train* dan *data test* sehingga model CNN-GRU lebih mampu melakukan generalisasi pada *data test*. Namun, CNN-LSTM unggul pada *data validation*, menunjukkan performa yang sedikit lebih baik dalam proses validasi. Secara keseluruhan, CNN-GRU dapat dianggap sebagai model yang menunjukkan hasil terbaik pada kasus ini, karena akurasi pada *data test* yang lebih tinggi. Hal ini dapat terjadi akibat kemampuan model CNN-GRU dalam mempelajari fitur penting pada data latih dapat menghasilkan *generalization* yang lebih baik pada *data test*. Selain itu, *data test* yang lebih mirip dengan *data train* dibandingkan *data validation*, dapat membuat model CNN-GRU bisa memiliki performa lebih baik di *data test*.

Conclusions / Kesimpulan

Penelitian ini mengeksplorasi metode pengenalan gestur tangan menggunakan dua pendekatan deep learning, yakni *Convolutional Neural Network* dengan *Long Short-Term Memory* (CNN-LSTM) dan *Convolutional Neural Network* dengan *Gated Recurrent Unit* (CNN-GRU), dengan menggunakan dataset video Jester. Melalui serangkaian eksperimen mendalam, ditemukan bahwa model CNN-GRU menunjukkan performa yang sedikit lebih baik dengan akurasi 76% pada data validation dan 70% pada data test, dibandingkan CNN-LSTM yang mencapai 82% pada *validation* dan 60% pada *test*. Meskipun kedua model memiliki kelebihan masing-masing, CNN-GRU dianggap lebih unggul karena kemampuan generalisasi yang lebih baik pada data test. Kajian mendalam menunjukkan bahwa kedua model memperlihatkan kinerja optimal pada gerakan spesifik seperti 'Mengetuk Jari' dan 'Menggeser Dua Jari Ke Bawah', namun menghadapi tantangan substantif dalam mengklasifikasikan gerakan 'Menjauhkan Dua Jari'. Penelitian ini juga mengungkap karakteristik penting terkait *overfitting* pada model, yang tercermin dari divergensi signifikan antara akurasi *training* dan validasi, mengindikasikan perlunya strategi refinement untuk meningkatkan generalisasi model. Untuk pengembangan penelitian selanjutnya, disarankan untuk melakukan penyempurnaan melalui optimalisasi *preprocessing* dengan teknik yang lebih canggih, seperti augmentasi data video, normalisasi yang lebih detail, dan seleksi fitur yang lebih selektif, serta eksperimentasi dengan berbagai *hyperparameter* seperti jumlah lapisan, ukuran filter *konvolusi, learning rate*, dan regularisasi untuk mengurangi *overfitting* dan meningkatkan generalisasi model dalam pengenalan gestur tangan.

References

[1] T. Mender, “20bn-jester Dataset,” Kaggle, https://www.kaggle.com/datasets/toxicmender/20bn-jester, 20 Des 2023..

[2] F. I. Hadinata dan S. A. Sanjaya, “BISINDO Sign Language Recognition : A Systematic Literature Review of Deep Learning Techniques for Image Processing”, *Indonesian Journal of Computer Science*, Vol. 12, No. 6, pp. 3281-3294, 20 Des 2023.

[3] M. M. Fresmanda, Istiadi dan S. W. Irianandа, “Deteksi Objek Video Bahasa Isyarat Untuk Anak Tuna Rungu Dan Tuna Wicara Menggunakan YOLOv8”, *Jurnal Komputer, Informasi dan Teknologi,* Vol. 4, No. 2, pp. 1-9, 1 Des 2024.

[4] G. B. Prananta, H. A. Azzikri dan C. Rozikin, “Deteksi Pengenalan Gestur Tangan Secara Real-Time Menggunakan Jaringan Saraf Tiruan Konvolusional”, Jurnal METHODIKA, Vol. 9, No. 2, pp. 30-34, 2 Sept 2023.

[5] Ran Bi, “Sensor-based gesture recognition with convolutional neural networks”, *Proceedings of the 3rd International Conference on Signal Processing and Machine Learning,* pp. 456-462, doi : 10.54254/2755-2721/4/2023305.

[6] J. Materzynska, G. Berger, I. Bax, and R. Memisevic, “The jester dataset: A large-scale video dataset of human gestures,” *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 2874–2882, Oct. 2019. doi:10.1109/iccvw.2019.00349

[7] Zhu, M., Zhang, C., Wang, J., Sun, L., dan Fu, M., "Robust Hand Gesture Recognition Using a Deformable Dual-Stream Fusion Network Based on CNN-TCN for FMCW Radar," **Sensors**, vol. 23, no. 20, pp. 8570, 19 Okt. 2023. doi: 10.3390/s23208570.

[8] Huang, Z., Yang, F., Xu, F., Song, X., dan Tsui, K.-L., "Convolutional Gated Recurrent Unit–Recurrent Neural Network for State-of-Charge Estimation of Lithium-Ion Batteries," **IEEE Access**, vol. 7, pp. 93139–93149, 10 Juli 2019. doi: 10.1109/ACCESS.2019.2928037.

[9] Roberto Interdonato, "Visual representation of the Gated Recurrent Unit cell: The GRU cell has two internal gates," ResearchGate, diakses pada 17 Juni 2024. Tersedia: doi:10.1109/iccwv.2019.00349

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia