

Identifikasi Alfabet Bahasa Isyarat Indonesia dengan Menggunakan Convolutional LSTM

*Andrian Dwi Baitur Rizky, Muhammad Aulia Faqihuddin,
Fajar Fatha Romadhan, Indah Agustien Siradjuddin*

*Teknik Informatika, Universitas Trunojoyo Madura
Jl. Raya Telang, PO BOX 2, Kecamatan Kamal, Kabupaten Bangkalan, Jawa Timur 69162
Email : 200411100210@student.trunojoyo.ac.id*

Abstrak. Bagi orang dengan gangguan pendengaran, bahasa isyarat Indonesia sangat penting dikarenakan mereka dapat berkomunikasi dengan menggunakan isyarat tangan yang menunjukkan arti tertentu. Namun, karena beberapa konsep tidak memiliki isyarat khusus, maka harus diuraikan menggunakan huruf alfabet. Biasanya seseorang akan merasa sulit ketika berkomunikasi dengan orang yang memiliki gangguan pendengaran dan tidak mengenal bahasa isyarat. Oleh karena itu, diperlukan suatu model yang dapat mengenali bahasa isyarat untuk membantu pemula yang belajar bahasa isyarat berbasis abjad. Tujuan dari penelitian ini adalah untuk mengembangkan model pengenalan bahasa isyarat Indonesia (BISINDO) dari data timeseries berupa video yang terdiri dari 26 kategori dengan menggunakan metode convolutional LSTM. Metode ini merupakan inovasi dalam dunia deep learning yang menggabungkan Long-Short Term Memory (LSTM) dan Convolutional Neural Network (CNN). Penelitian ini menggunakan model convolutional LSTM 2D dengan arsitektur yang sudah direncanakan. Hasil menunjukkan bahwa model memiliki tingkat akurasi model sekitar 68%, menggambarkan kemampuan yang signifikan dalam memprediksi gerakan alfabet dalam Bahasa Isyarat Indonesia (BISINDO). Namun, perlu ketahui bahwa terdapat 32% kemungkinan prediksi yang tidak tepat, perlunya penelitian yang lebih mendalam untuk mengatasi kerumitan variasi gerakan dalam bahasa isyarat.

Katakunci: BISINDO, LSTM, CNN, Bahasa isyarat, Data timeseries.

1. Pendahuluan

Bagi penyandang tunarungu, interaksi sosial memiliki peran yang sangat penting untuk memahami satu sama lain dalam kehidupan masyarakat [1]. Mengidentifikasi isyarat tangan juga menjadi hal yang sangat penting juga untuk mengatasi berbagai tantangan dan mempermudah kehidupan mereka, terutama bagi penyandang tunarungu [2]. Dalam konteks komunikasi, bahasa isyarat adalah bentuk gerakan tangan terstruktur yang melibatkan gerakan visual dan penggunaan berbagai bagian tubuh seperti jari, tangan, lengan, kepala, tubuh, dan ekspresi wajah. Bahasa isyarat menjadi alat penting untuk berinteraksi sehari-hari bagi penyandang tunarungu dan tuna wicara [3]. Namun, sedikit orang di luar komunitas penyandang tunarungu yang mengerti bahasa isyarat, dan hanya sedikit yang benar-benar memahaminya. Hal ini membuat hambatan dalam komunikasi antara komunitas penyandang tunarungu dengan masyarakat umum, yang belum sepenuhnya memahami bahwa bahasa isyarat harus dieja dengan menggunakan isyarat untuk setiap huruf [4]. Dengan demikian, penting untuk terus mempromosikan pemahaman dan inklusi mengenai bahasa isyarat di kalangan masyarakat umum, sehingga memfasilitasi komunikasi yang lebih efektif dan inklusif bagi semua orang.

Pemerintah Indonesia resmi mengakui Sistem Bahasa Isyarat Indonesia (SIBI) sebagai bahasa isyarat resmi di Sekolah Luar Biasa (SLB) sejak tahun 1996. Akan tetapi, masyarakat difabel di Indonesia cenderung menggunakan Bahasa Isyarat Indonesia (BISINDO). Alasannya adalah karena BISINDO berasal dari Bahasa Indonesia yang merupakan bahasa ibu yang mereka gunakan sehari-hari. Sehingga, satu gerakan isyarat dalam Bisindo dapat merepresentasikan kata-kata yang juga dikenal oleh orang yang tidak memiliki gangguan pendengaran. Di sisi lain, gerakan isyarat dalam SIBI mengikuti tata bahasa yang digunakan oleh individu. Hal ini kadang-kadang dapat menyebabkan tidak mencerminkan ekspresi yang sebenarnya dari apa yang didengar oleh semua orang. Komunitas penyandang disabilitas pendengaran telah mendesak pemerintah sejak tahun 1975 untuk mengakui BISINDO sebagai bahasa

resmi. Permintaan ini terus diupayakan karena penggunaan BISINDO memudahkan komunikasi bagi penyandang disabilitas pendengaran, terutama dalam menyampaikan aspirasi mereka kepada orang-orang yang tidak memiliki gangguan pendengaran [5].

Karena itu, tujuan dari penelitian ini adalah mengembangkan model pengenalan bahasa isyarat BISINDO dengan menggunakan metode *Deep Learning*. Pemilihan metode *Deep Learning* diambil karena popularitasnya di dunia ilmu komputer dan keberhasilannya terbukti dalam mengklasifikasikan data dari kumpulan video. Salah satu hal baru dari penelitian ini adalah penerapan langkah-langkah pra-pemrosesan pada video selama tahap pelatihan dan pengujian untuk meningkatkan performa model.

2. Pembahasan

Konvolusi merupakan sebuah proses yang digunakan untuk mengekstrak fitur dari data spasial seperti pada citra. Konvolusi bekerja dengan melibatkan kernel atau filter konvolusi. Kernel akan dikalikan menggunakan perkalian titik dan digeser secara berurutan melintasi seluruh citra *input*. Hasil perkalian titik tersebut dijumlahkan sehingga menghasilkan nilai tunggal dan menghasilkan *feature map*. Pemanfaatan operasi konvolusi dalam *neural network* dapat dilihat pada CNN. CNN atau *Convolutional Neural Network* merupakan arsitektur jaringan syaraf tiruan yang memanfaatkan layer konvolusi. CNN banyak digunakan sebagai layer yang digunakan untuk mengekstrak fitur dari data inputan seperti citra. Layer konvolusi pada CNN dinilai dapat mengurangi beban komputasi dari *neural network*[6].

Recurrent Neural Network (RNN) merupakan suatu metode dalam *machine learning* yang digunakan untuk memproses data sekuensial atau *time series*. Pada data yang berbentuk sekuensial atau *time series*, nilai data $t+1$ akan dipengaruhi oleh data $t-0$ dan $t-1$. RNN memiliki kemampuan untuk mengingat input sebelumnya dengan menggunakan konsep memori internal atau *state* yang memungkinkan informasi sebelumnya dapat mempengaruhi pemrosesan informasi saat ini. RNN secara teori dapat melihat *long term dependency* jauh-jauh dari masa lampau, sehingga semua *input* dari masa lampau dapat memengaruhi *output*. Tapi pada implementasinya RNN memiliki masalah dengan *input* yang terlalu jauh di masa lampau. Untuk mengatasi hal tersebut arsitektur RNN dikembangkan sehingga tercipta arsitektur yang dikenal sebagai *Long Short Term Memory* (LSTM).

LSTM merupakan salah satu bentuk RNN yang dapat mengatasi masalah korelasi deret waktu baik dalam waktu singkat maupun lama dengan menggunakan lapisan tersembunyi sebagai sel memori. Ciri khas dari LSTM yaitu selain ada *hidden state* yang diberikan ke time step berikutnya terdapat juga *cell state*. *Cell state* ini seperti state internal pada LSTM. *Cell state* ini akan membawa informasi dari masa lampau. Informasi tersebut dapat dilupakan atau dibuang atau dapat ditambahkan pada *time step* berikutnya sehingga dapat diproses di *hidden state*. LSTM sebagai pengembangan dari RNN telah terbukti stabil dan kuat dalam memodelkan ketergantungan jarak jauh dalam berbagai studi sebelumnya[7].

Meski demikian, LSTM tidak dapat menerima *input* data spasial seperti citra. Arsitektur LSTM dirancang untuk memproses data yang berurutan seperti suara, teks, atau data *time series*. Berbeda dengan citra yang pada dasarnya merupakan data spasial dengan dimensi yang lebih besar

daripada data yang berurutan. Data video merupakan data berurutan yang terdiri atas banyak *frame* yang berupa citra. Untuk memproses data video, arsitektur LSTM dapat dikombinasikan dengan CNN. Arsitektur ini dinamakan *Convolutional LSTM*. Berbeda dengan LSTM, *Convolutional LSTM* dapat menerima *input* berupa citra dan menyimpan informasi spasialnya sehingga memfasilitasi rekonstruksi data[8]. Dalam *Convolutional LSTM*, layer konvolusi digunakan untuk mengekstrak fitur spasial dari citra. Sedangkan LSTM digunakan untuk mengolah dan memahami urutan atau deret waktu pada video.

2.1. Dataset

Dataset yang digunakan dalam penelitian ini berupa kumpulan video yang memperagakan gerakan alfabet pada bahasa isyarat Indonesia (BISINDO) dari huruf A – Z. Dataset dapat diakses pada laman *trunojoyo-ai.id*[9]. Untuk masing – masing alfabet terdapat 15 hingga 73 video. Sebelum memasuki pelatihan, dilakukan praproses dataset terlebih dahulu. Peneliti mencoba untuk mengambil 25 frame dari setiap video, dan mengatur ukuran gambar untuk setiap frame sebesar 100x100 piksel.

2.2. Parameter Model

Model pada penelitian ini dibentuk dari 4 layer Convolutional LSTM 2D dengan parameter , Max Pooling, dan Time Distributed, serta 1 layer flatten dan 1 layer dense.

Untuk arsitektur yang lebih detail dapat dilihat pada tabel 1:

Tabel 1. Arsitektur Model

No	Layer	Parameter	Output Shape
1	Convolutional LSTM 2D	Filters : 4; Kernel Size : (3,3); Activation : tanh; Data Format : channel_last; Recurrent Dropout : 0,2; Return Sequences : True; Input Shape : (50, 100, 100, 3)	(None, 50, 98, 98, 4)
2	Max Pooling 3D	Pool size : (1, 2, 2); Padding : 'same'; Data format : 'channels_last'	(None, 50, 49, 49, 4)
3	Time Distributed	Dropout : 0,2	(None, 50, 49, 49, 4)
4	Convolutional LSTM 2D	Filters : 8; Kernel Size : (3,3); Activation : tanh; Data Format : channel_last; Recurrent Dropout : 0,2; Return Sequences : True	(None, 50, 47, 47, 8)
5	Max Pooling 3D	Pool size : (1, 2, 2); Padding : 'same'; Data format : 'channels_last'	(None, 50, 24, 24, 8)
6	Time Distributed	Dropout : 0,2	(None, 50, 24, 24, 8)
7	Convolutional LSTM 2D	Filters : 14; Kernel Size : (3,3);	(None, 50, 22, 22, 14)

No	Layer	Parameter	Output Shape
		Activation : tanh; Data Format : channel_last; Recurrent Dropout : 0,2; Return Sequences : True;	
8	Max Pooling 3D	Pool size : (1, 2, 2); Padding : 'same'; Data format : 'channels_last'	(None, 50, 11, 11, 14)
9	Time Distributed	Dropout : 0,2	(None, 50, 11, 11, 14)
10	Convolutional LSTM 2D	Filters : 16; Kernel Size : (3,3); Activation : tanh; Data Format : channel_last; Recurrent Dropout : 0,2; Return Sequences : True;	(None, 50, 9, 9, 16)
11	Max Pooling 3D	Pool size : (1, 2, 2); Padding : 'same'; Data format : 'channels_last'	(None, 50, 5, 5, 16)
12	Time Distributed	Dropout : 0,2	(None, 50, 5, 5, 16)
13	Flatten		(None, 20000)
14	Dense	Sesuai jumlah class (26)	(None, 26)

2.3. Hasil Ujicoba

Proses pelatihan model dilakukan sebanyak 100 epoch, menggunakan optimizer adam sehingga dengan learning rate default sebesar 0,001. Sebanyak 80% data digunakan untuk data training, dan 25% dari data training digunakan untuk data testing. Dari ujicoba yang telah dilakukan didapatkan nilai akurasi model sebesar 68%.

Tabel 2 Detail Pembagian Data Video

Huruf	Jumlah Video	Jumlah Video (Training)	Jumlah Video (Testing)
A	40	32	9
B	73	58	8
C	15	12	3
D	16	13	5
E	15	12	4
F	26	21	6
G	20	16	3
H	20	16	5
I	20	16	4
J	22	18	6
K	25	20	3
L	15	12	1
M	15	12	1

Huruf	Jumlah Video	Jumlah Video (Training)	Jumlah Video (Testing)
N	15	12	4
O	15	12	4
P	15	12	4
Q	15	12	1
R	15	12	8
S	15	12	4
T	15	12	5
U	15	12	2
V	15	12	5
W	15	12	1
X	15	12	2
Y	15	12	3
Z	15	12	3
TOTAL	517	414	104

Pada tabel 2 ditunjukkan detail pembagian dataset untuk proses pelatihan model. Total seluruh data adalah 517 video, kemudian jumlah data yang digunakan untuk proses pelatihan sebanyak 80% dari total data yaitu 414 video, dan jumlah data yang digunakan untuk proses testing sebanyak 25% dari total data pelatihan yaitu 104 video.



Gambar 1. Contoh Data Alfabet "C"



Gambar 2 Contoh Data Alfabet "G"

Pada Gambar 1 dan Gambar 2 ditunjukkan contoh hasil prediksi video dari kategori alfabet “C” dan “G”. Hal tersebut menunjukkan bahwa model dapat mendeteksi video gerakan alfabet bahasa isyarat tangan Indonesia dengan baik.

Kemudian peneliti mencoba memprediksi kembali data ketiga, yaitu huruf T. Berikut adalah gambaran dari video data T yang akan dilakukan deteksi oleh model:



Gambar 3 Contoh Data Alfabet "T"

Pada Gambar 3 ditunjukkan contoh hasil prediksi untuk huruf “T”, model mendeteksi bahwa video tersebut merupakan gerakan untuk huruf “B”. Hal ini menunjukkan bahwa model tidak sepenuhnya benar dalam memprediksi gerakan pada video yang diberikan dikarenakan akurasi yang didapatkan oleh model hanya sebesar 68% yang artinya ada 32% kemungkinan model akan salah dalam memprediksi data.

Nilai akurasi model didapatkan dengan menggunakan rumus *accuracy score* yaitu:

$$\begin{aligned}
 \text{Akurasi} &= \frac{\text{Banyaknya prediksi yang benar}}{\text{Total prediksi yang dilakukan}} \\
 &= \frac{TP + TN}{TP + FP + TN + FN} \quad (1)
 \end{aligned}$$

dimana:

TP : True Positive
FP : False Positive
TN : True Negative
FN : False Negative

Tabel 3 Confusion Matrix

Huruf	TP	FP	FN	TN
A	9	1	0	94
B	8	0	0	96
C	3	1	0	100
D	3	0	2	99
E	3	1	1	99
F	4	1	2	97
G	2	3	1	98
H	1	1	4	98
I	4	3	0	97
J	6	0	0	98
K	3	0	0	101

Huruf	TP	FP	FN	TN
L	1	0	0	103
M	1	2	0	101
N	2	1	2	99
O	3	2	1	98
P	1	2	3	98
Q	0	1	1	102
R	2	1	6	95
S	1	1	3	99
T	3	1	2	98
U	2	5	0	97
V	1	1	4	98
W	1	1	0	102
X	2	0	0	102
Y	3	2	0	99
Z	3	1	0	100
TOTAL	72	32	32	2568

Pada tabel 3 disajikan tabel confusion matrix yang mencerminkan hasil evaluasi kinerja prediksi untuk huruf A sampai Z. Tabel tersebut memberikan gambaran yang jelas tentang sejauh mana sistem kami mampu mengklasifikasikan data dengan benar dan diidentifikasi dengan benar. Dapat dilihat total dari matrix untuk setiap huruf adalah 104, hal ini sesuai dengan banyaknya data testing yaitu sebanyak 104 video. Penggunaan confusion matrix ini memungkinkan kita untuk mengukur kinerja prediksi dengan lebih rinci. Dengan memeriksa metrik-metrik seperti akurasi, presisi, dan recall untuk setiap kelas, kita dapat memahami lebih baik dimana sistem berhasil dan dimana mungkin masih ada kekurangan.

Tabel confusion matrix ini menggambarkan elemen-elemen berikut:

- True Positives (TP) : Jumlah data yang diprediksi secara tepat sebagai kelas positif
- False Positives (FP) : Jumlah data yang diprediksi secara tidak tepat sebagai kelas positif.
- False Negatives (FN) : Jumlah data yang diprediksi secara tidak tepat sebagai kelas negatif.
- True Negatives (TN) : Jumlah data yang diprediksi secara tepat sebagai kelas negatif.

Total dari masing-masing elemen ini memberikan gambaran umum tentang performa sistem klasifikasi. Dengan demikian, confusion matrix membantu kita dalam mengidentifikasi area-area yang memerlukan perbaikan, serta menilai efektivitas sistem klasifikasi secara keseluruhan.

3. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan dapat disimpulkan bahwa arsitektur Convolutional LSTM yang dibangun cukup baik untuk digunakan dalam melakukan pelatihan model dengan menggunakan data berbentuk video atau deretan beberapa gambar. Dari uji coba yang telah dilakukan, peneliti berhasil mendapatkan nilai akurasi model sebesar 68%. Model tersebut telah dapat memprediksi data baru yang dengan baik, namun belum dapat dikatakan optimal dikarenakan ada 32% kemungkinan model akan salah dalam memprediksi data.

Ucapan Terima Kasih

Terima kasih peneliti sampaikan kepada Lembaga Penelitian dan Pengabdian kepada Masyarakat dalam skema Hibah Grup Riset dengan No Kontrak Penelitian 5714/UN46.4.1/PT.01.03/2023

Daftar Pustaka

- [1] L. K. S. Tolentino, R. O. Serfa Juan, A. C. Thio-ac, M. A. B. Pamahoy, J. R. R. Forteza, and X. J. O. Garcia, "Static sign language recognition using deep learning," *International Journal of Machine Learning and Computing*, vol. 9, no. 6, pp. 821–827, 2019, doi: 10.18178/ijmlc.2019.9.6.879.
- [2] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 1, pp. 131–153, 2019, doi: 10.1007/s13042-017-0705-5.
- [3] A. S. M. Miah, J. Shin, M. A. M. Hasan, and M. A. Rahim, "BenSignNet: Bengali Sign Language Alphabet Recognition Using Concatenated Segmentation and Convolutional Neural Network," *Applied Sciences (Switzerland)*, vol. 12, no. 8, 2022, doi: 10.3390/app12083933.
- [4] R. Rastgoo, K. Kiani, and S. Escalera, "Hand sign language recognition using multi-view hand skeleton," *Expert Systems with Applications*, vol. 150, p. 113336, 2020, doi: 10.1016/j.eswa.2020.113336.
- [5] M. Susanty, R. Z. Fadillah, and A. Irawan, "Model Penerjemah Bahasa Isyarat Indonesia (BISINDO) Menggunakan Pendekatan Transfer Learning," *Petir*, vol. 15, no. 1, pp. 1–9, 2021, doi: 10.33322/petir.v15i1.1289.
- [6] Y. Widhiyasana, T. Semiawan, I. Gibran, A. Mudzakir, and M. R. Noor, "Penerapan Convolutional Long Short-Term Memory untuk Klasifikasi Teks Berita Bahasa Indonesia (Convolutional Long Short-Term Memory Implementation for Indonesian News Classification)," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi* |, vol. 10, no. 4, 2021.
- [7] X. Shi *et al.*, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting."
- [8] Luo Weixin, Liu Wen, and Gao Shenghua, "REMEMBERING HISTORY WITH CONVOLUTIONAL LSTM FOR ANOMALY DETECTION," in *International Conference on Multimedia and Expo*, 2017.
- [9] Trunojoyo AI, "Bisindo Dataset ," <https://trunojoyo-ai.id/dataset/bisindo.html>.