PERBANDINGAN MODEL CNN DAN LSTM DALAM PENGENALAN PERINTAH SUARA PADA ROBOT MINI FORKLIFT

M. F. Aulian[1]*, C. P. Tresnawan[2], T. R. Datuela[3], dan Irmawan[4]

[1]Teknik Elektro, Universitas Sriwijaya, Palembang
[2]Teknik Elektro, Universitas Sriwijaya, Palembang
[3]Teknik Elektro, Universitas Sriwijaya, Palembang
[4]Teknik Elektro, Universitas Sriwijaya, Palembang
[*]*Corresponding author* e-mail: fikri141589@gmail.com

ABSTRAK: Interaksi antara manusia dan mesin menjadi kunci dalam optimalisasi kegiatan operasional memindahkan barang di lingkungan industri seperti gudang. Sistem kontrol menggunakan suara memerlukan performa yang sangat bergantung pada keandalan model pengenalan suara yang digunakan. Penelitian ini bertujuan untuk mengembangkan dan menganalisis performa dua arsitektur deep learning, *Long Short-Term Memory* (LSTM) dan *Convolutional Neural Network* (CNN), untuk sistem kontrol suara real-time pada prototipe robot forklift. Untuk meningkatkan kemampuan model menghadapi variasi suara, diterapkan teknik augmentasi data berupa *Gaussian Noise* dan *Time Shift Spectrogram* selama proses pelatihan. Hasil evaluasi menunjukkan bahwa model CNN mencapai akurasi training yang lebih tinggi sebesar 93.6% (loss: 0.1857) dibandingkan LSTM dengan akurasi 89.4% (loss: 0.307). Namun, pada tahap validasi, model LSTM menunjukkan kemampuan yang sedikit lebih baik dengan akurasi validasi 83.8% (val_loss: 0.451), sementara CNN mencatatkan akurasi validasi 82.05% (val_loss: 0.732). Pengujian real-time pada prototipe dengan beberapa penguji bersuara berbeda dan gender yang berbeda mengonfirmasi bahwa kedua model mampu mengeksekusi perintah yang diucapkan dalam kalimat panjang maupun pendek. Selain itu, ditemukan bahwa kebisingan lingkungan (ambient noise) terkadang dapat menyebabkan kegagalan deteksi, yang menyoroti tantangan utama dalam implementasi praktis sistem di lapangan.

Kata Kunci: Pengenalan Suara, CNN, LSTM, Spektogram, Perintah Suara.

ABSTRACT: *Human-machine interaction is key to optimizing operational activities for moving goods in industrial environments such as warehouses. Voice-based control systems require performance that is highly dependent on the reliability of the speech recognition model used. This study aims to develop and analyze the performance of two deep learning architectures, Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN), for a real-time voice control system on a forklift robot prototype. To improve the model's ability to handle sound variations, data augmentation techniques such as Gaussian Noise and Time Shift Spectrogram were applied during the training process. The evaluation results show that the CNN model achieved a higher training accuracy of 93.6% (loss: 0.1857) compared to the LSTM with an accuracy of 89.4% (loss: 0.307). However, in the validation stage, the LSTM model showed slightly better performance with a validation accuracy of 83.8% (val_loss: 0.451), while the CNN recorded a validation accuracy of 82.05% (val_loss: 0.732). Real-time testing on prototypes with multiple testers with different voices and genders confirmed that both models were capable of executing spoken commands in both long and short sentences. Furthermore, it was found that ambient noise can sometimes cause detection failures, highlighting a major challenge in the practical implementation of the system in the field.*

Keywords: Speech Recognition, CNN, LSTM, Spectrogram, Voice Command.

1 Pendahuluan

Interaksi Manusia dan Robot saat ini telah menjadi komponen krusial dalam perkembangan industri, khususnya di sektor logistik dan manufaktur. Robot forklift, baik dalam skala industri maupun sebagai prototipe mini untuk penelitian, membutuhkan metode kontrol yang efisien, intuitif, dan aman. Kontrol berbasis perintah suara (voice command) memberikan pengalaman dalam interaksi yang mudah dibandingkan metode tradisional (seperti joystick atau tombol), karena mengurangi beban kognitif operator, dan meningkatkan fleksibilitas operasional. Implementasi sistem ini, yang menjembatani Natural Language Understanding (NLU) dengan aktuasi robotik, bergantung pada kemampuan Speech Recognition yang akurat [1].

Namun, tantangan utama dalam mengimplementasikan sistem ini adalah akurasi pengenalan suara dan kemampuan adaptasi terhadap berbagai kondisi lingkungan, seperti kebisingan latar belakang atau variasi suara manusia. Deep learning, dengan kemampuannya untuk memproses data dalam jumlah besar dan mempelajari fitur-fitur kompleks, menawarkan solusi yang menjanjikan untuk mengatasi tantangan ini [2]. Selain itu, penggunaan spectrogram sebagai input untuk model deep learning telah menunjukkan hasil yang sangat baik dalam meningkatkan akurasi sistem pengenalan suara. Hal ini disebabkan oleh kemampuan spectrogram untuk menyajikan representasi visual sinyal audio dalam domain waktu dan frekuensi secara komprehensif, sehingga model dapat dengan lebih efektif mengidentifikasi pola-pola penting serta fitur-fitur unik dari sinyal suara. Selain itu, teknik pra proses menggunakan spectrogram juga membantu dalam mengurangi gangguan noise dan menonjolkan karakteristik utama yang relevan, sehingga lebih informatif dan dapat digunakan untuk tugas-tugas pengenalan suara [3].

Dua arsitektur yang dominan dalam pengenalan suara adalah Convolutional Neural Networks (CNN) dan Long Short-Term Memory (LSTM) yang biasanya dirancang untuk visi komputer, terbukti efektif dalam mengekstraksi fitur lokal-spasial yang hierarkis dari representasi audio (seperti spectrogram), sehingga unggul dalam mengidentifikasi fonem atau kata kunci yang terisolasi. Di sisi lain, LSTM, sebagai varian dari Recurrent Neural Networks (RNN), dirancang secara eksplisit untuk memodelkan dependensi temporal dan konteks jangka panjang, yang secara teoritis untuk memahami sebuah kalimat atau perintah utuh [4].

Dengan demikian, integrasi robot pengangkut barang dengan kontrol suara berbasis deep learning tidak hanya meningkatkan efisiensi operasional, tetapi juga membuka peluang baru untuk otomatisasi di berbagai bidang. Penelitian lebih lanjut diperlukan untuk mengoptimalkan performa sistem ini, terutama dalam hal akurasi pengenalan suara dan adaptasi terhadap lingkungan yang dinamis.

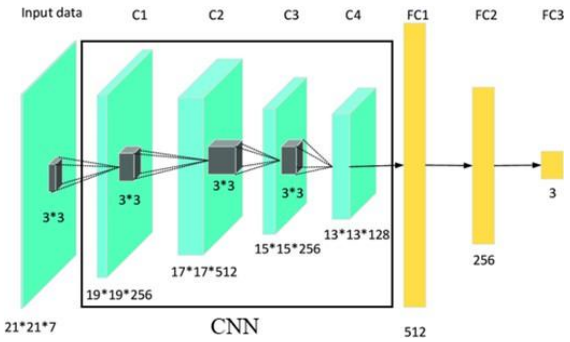
2 Dasar Teori

Pengenalan suara berbasis deep learning merupakan teknologi yang memungkinkan komputer atau robot untuk memahami perintah verbal manusia dan mengkonversinya menjadi tindakan. Contohnya suara digunakan untuk mengendalikan robot melalui perintah verbal seperti "maju", "mundur", "kanan", "kiri", dan lainnya. Suara juga memainkan peran penting dalam interaksi manusia-mesin, termasuk dalam asisten virtual (voice assistants) [5].

2.1 *Convolutional Neural Network* (CNN)

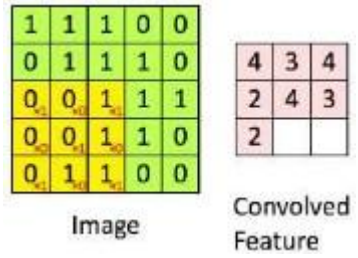
Convolutional Neural Network (CNN) adalah tumpukan modul pembelajaran berlapis yang sangat cocok untuk menangani dataset dua dimensi (misalnya, gambar). CNN adalah subkelas jaringan saraf tiruan yang menggabungkan pemrosesan nonlinier neuron lapisan tersembunyi dengan sifat-sifat penting pembagian bobot (atas sub-gambar yang dapat disesuaikan—disebut filter konvolusional), pengumpulan, dan pengambilan sampel ulang [6].

Representasi spektrogram dalam suatu data sinyal suara menunjukkan korelasi yang sangat kuat dalam waktu dan frekuensi. Karena karakteristik spektrogram merupakan input yang sesuai untuk alur pemrosesan CNN yang mengharuskan pelestarian lokalitas pada sumbu frekuensi dan waktu. Untuk sinyal suara, pemodelan korelasi lokal dengan CNN akan bermanfaat. CNN juga dapat secara efektif mengekstraksi fitur struktural dari spektrogram dan mengurangi kompleksitas model melalui pembagian bobot [2].



Gambar 1. Lapisan Convolutional Neural Network

pada CNN, data yang dipropagasikan pada lapisan jaringan adalah data dua dimensi, sehingga operasi linear dan parameter bobot pada CNN berbeda [7]. Pada CNN operasi linear menggunakan operasi konvolusi dan nilai bobot tidak lagi satu dimensi.



Gambar 2. Operasi Konvolusi

Pooling layer adalah komponen dalam CNN yang berfungsi untuk mengurasi dimensi fitur sambil mempertahankan informasi penting dari gambar atau data input. Pooling layer biasanya ditempatkan setelah convolutional layer untuk menurunkan kompleksitas komputasi dan meningkatkan generalisasi model. Adapun fungsi dari pooling itu sendiri adalah untuk mengurasi parameter dalam jaringan, mempercepat proses pelatihan dan menurunkan ukuran data, dan mengurangi sensitivitas terhadap translasi kecil dalam gambar.

2.2 *Long Short-Term Memory* (LSTM)

Long short-term memory (LSTM) adalah jenis *Recurrent Neural Network* yang dirancang khusus untuk mencegah keluaran jaringan syaraf untuk input tertentu menurun atau meledak saat berputar melalui umpan balik. Umpan balik inilah yang memungkinkan jaringan recurrent menjadi lebih baik dalam pengenalan pola dibandingkan dengan jaringan syaraf lainnya. Memori dari input yang telah ada sangat penting untuk menyelesaikan tugas pembelajaran urutan, dan jaringan Memori Jangka Pendek yang Panjang memberikan kinerja yang lebih baik dibandingkan dengan arsitektur RNN lainnya dengan mengurangi apa yang disebut masalah gradien menghilang [8].

3 Speech Recognition dan Arsitektur

3.1 Dataset

Kinerja dan kemampuan suatu model dari sebuah arsitektur deep learning, baik CNN maupun LSTM, ditentukan oleh kualitas dan representasi data latih. Untuk memastikan validitas komparasi dan relevansi hasil penelitian dengan aplikasi di dunia nyata (kontrol robot via HP), sebuah dataset yang representatif mutlak diperlukan. Oleh karena itu, sebuah dataset perintah suara kustom telah dikumpulkan sebanyak 6.318 sampel dengan variasi gender dan besar kecil suara untuk menyebutkan kalimat perintah yang panjang dan pendek seperti tabel 1. Berikut adalah list dataset perintah suara :

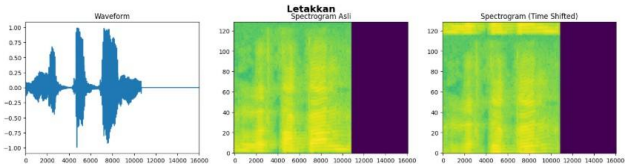
Tabel 1. List perintah suara.

No	Perintah
1	Ikut saya
2	Stop
3	Berhenti
4	Ke sini
5	Kembali ke awal
6	Bawa ke lokasi A
7	Bawa ke lokasi B
8	Pelankan
9	Cepat
10	Tunggu
11	Ambilkan barang di tempat A
12	Ambilkan barang di tempat B
13	Maju
14	Angkat
15	Letakkan

3.2 Preprocessing

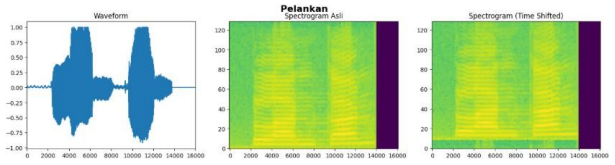
Data audio mentah dalam dataset berbentuk waveform, data ini merepresentasikan sinyal 1-dimensi (amplitudo terhadap waktu). Secara khusus, arsitektur Convolutional Neural Networks (CNN) dirancang untuk mengekstraksi fitur spasial hirarkis dari data 2D (seperti gambar).

Untuk memanfaatkan kemampuan CNN ini, terlebih dahulu dilakukan preprocessing setiap sinyal waveform 1D menjadi representasi visual 2D, yaitu spektrogram. Proses ini menggunakan Short-Time Fourier Transform (STFT) dengan mengubah data audio menjadi "gambar" yang memetakan waktu pada sumbu-X, frekuensi pada sumbu-Y, dan amplitudo (energi) sebagai intensitas. Hal ini memungkinkan model CNN untuk "melihat" pola-pola vokal dan konsonan sebagai fitur visual yang dapat dipelajari.



(a) (b) (c)

Gambar 3. pemrosesan suara perintah “Letakkan” a) waveform, b) spektrogram, c) spektrogram time shifting (augmentasi)



(a) (b) (c)

Gambar 4. pemrosesan suara perintah “Pelankan” a) waveform, b) spektrogram, c) spektrogram time shifting (augmentasi)

Untuk mengatasi variasi dari berbagai suara dilakukan augmentasi data yang sekaligus memperkaya jumlah data latih, teknik yang diterapkan teknik ***time-shifting*** pada spektrogram yang telah dihasilkan dapat dilihat pada gambar 3 dan gambar 4. Dengan menggeser spektrogram secara acak beberapa milidetik ke kiri atau ke kanan di sepanjang sumbu waktu.

4 Uji Coba dan Evaluasi

Uji coba dilakukan masing-masing 6 orang penguji dan penyebutan 15 perintah sehingga total percobaan sebanyak 90 data berbeda. Berikut adalah hasil percobaan yang menunjukkan keberhasilan model dalam menerima suara yang beragam dari perbedaan gender, perbedaan pitch nada, dan suara penguji yang belum diambil sampel suaranya

4.1 Model CNN

NO	Perintah	keberhasilan					
		Aditya	Ari	Reyhan	Rizal	Fikri	Bernika
1	Ikut Saya	1	1	0	1	0	0
2	Stop	1	1	1	0	0	1
3	Berhenti	1	0	0	0	0	1
4	Ke sini	1	1	1	1	1	1
5	Kembali ke awal	0	1	1	1	1	1
6	Bawa ke lokasi A	0	0	0	1	1	1
7	Bawa ke lokasi B	0	1	1	0	0	0
8	Pelankan	1	1	0	1	1	1
9	Cepat	1	1	1	1	1	1
10	Tunggu	0	1	1	0	0	1
11	Ambilkan barang di tempat A	0	0	0	0	0	0
12	Ambilkan barang di tempat B	0	0	1	0	0	1
13	Maju	0	0	1	1	1	1
14	Angkat	0	1	0	1	0	1
15	Letakkan	0	1	1	1	1	1
		Jumlah			Persentase		
	benar	53			58.89		
	salah	37			41.11		
		Model Training					
		accuracy: 0.9291 - loss: 0.1999 - val_accuracy: 0.8205 - val_loss: 0.7328					

Gambar 5. Hasil Percobaan model CNN (benar = 1; salah=0)

4.2 Model CNN

NO	Perintah	keberhasilan					
		Aditya	Ari	Reyhan	Rizal	Fikri	Bernika
1	Ikut Saya	1	1	1	1	1	1
2	Stop	1	1	1	0	1	1
3	Berhenti	1	1	1	1	1	1
4	Ke sini	1	1	1	1	1	1
5	Kembali ke awal	1	1	1	1	1	1
6	Bawa ke lokasi A	1	0	1	1	1	1
7	Bawa ke lokasi B	1	1	0	1	1	1
8	Pelankan	1	1	1	1	1	1
9	Cepat	1	1	1	1	0	1
10	Tunggu	1	1	1	1	0	1
11	Ambilkan barang di tempat A	0	0	0	1	0	0
12	Ambilkan barang di tempat B	0	0	0	1	0	0
13	Maju	1	1	0	1	1	1
14	Angkat	1	0	1	1	1	1
15	Letakkan	1	1	1	1	0	1
		Jumlah			Persentase		
	benar	72			80.00		
	salah	18			20.00		
		Model Training					
		accuracy: 0.8944 - loss: 0.3072 - val_accuracy: 0.8380 - val_loss: 0.4513					

Gambar 5. Hasil Percobaan model LSTM (benar = 1; salah=0)

4.3 Evaluasi

Dengan perbedaan hasil training untuk CNN dan LSTM yang berbeda signifikan pada validasi loss sebesar 0.738 (CNN) dengan 0.4513 (LSTM) menghasilkan perbedaan kemampuan dalam memprediksi kata yang diucapkan.

5 Kesimpulan

Penelitian ini telah melakukan analisis komparatif yang mendalam terhadap kinerja arsitektur *Convolutional Neural Networks* (CNN) murni dan *Long Short-Term Memory* (LSTM) murni untuk tugas klasifikasi perintah

suara real-time pada prototipe robot forklift mini. pengujian pada kalimat perintah yang pendek maupun kalimat panjang menunjukkan bahwa dibutuhkan variasi ekstra kepada perintah-perintah berkalimat panjang yang mempengaruhi keandalan robot dalam menjalankan tugas.

Daftar Pustaka

[1] B. Kuljić, S. János, dan S. Tibor, "Mobile robot controlled by voice," 5th *International Symposium on Intelligent Systems and Informatics*, SISY 2007, pp. 189–192, 2007, doi: 10.1109/SISY.2007.4342649.

[2] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, dan S. Poria, *A review of deep learning techniques for speech processing*, vol. 99. 2023. doi: 10.1016/j.inffus.2023.101869.

[3] M. A. Ismail, M. Eng. Dr. Ir. Djoko Purwanto, dan M. Sc. Fajar Budiman, ST., "Sistem Pengenalan Suara Untuk Perintah Pada Robot Sepak Bola Beroda," Institut Teknologi Sepuluh Nopember, 2019.

[4] J. Oruh, S. Viriri, dan A. Adegun, "Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition," IEEE Access, vol. 10, pp. 30069–30079, 2022, doi: 10.1109/ACCESS.2022.3159339.

[5] M. A. Suhendra, T. F. Parlaungan, dan T. Sumardi, "Pengenalan Suara sebagai Pengendali Mobile Robot dengan Metode Adaptive Neuro-Fuzzy Inference System," *TIME in Physics*, vol. 1, no. 1, pp. 43–49, Feb. 2023, doi: 10.11594/timeinphys.2023.v1i1p43-49.

[6] Eric Tatulli dan Thomas Hueber, "Feature Extraction Using Multimodal Convolutional Neural Networks For Visual Speech Recognition," 2017.

[7] W. S. E. Putra, "Klasifikasi Citra Menggunakan Convolutional Neural Network (CNN) pada Caltech 101," *JURNAL TEKNIK ITS* Vol., vol. 5, no. 1, Mar. 2016.

[8] J. Oruh, S. Viriri, dan A. Adegun, "Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition," *IEEE Access*, vol. 10, pp. 30069–30079, 2022, doi: 10.1109/ACCESS.2022.3159339.