BAB II

LANDASAN TEORI

2.1 Landasan Teori

2.1.1 Huruf Hiragana

Hiragana merupakan sistem penulisan asli dalam bahasa Jepang yang digunakan untuk menuliskan kata-kata sehari-hari dan tata bahasa. Hiragana sebenarnya berasal dari bentuk penyederhanaan huruf-huruf Kanji, dengan bentuk yang didominasi oleh garis-garis melengkung dan sederhana, yang memudahkan proses penulisan (Nurcholis dkk., 2021). Penyederhanaan ini dirancang agar lebih mudah digunakan dalam penulisan bahasa Jepang sehari-hari. Sistem penulisan ini terdiri dari 46 karakter, yang masing-masing mewakili satu bunyi tertentu dalam bahasa Jepang (Zhelita & Arni, 2023).

2.1.2 *Speech Recognition*

Speech recognition adalah teknologi yang mengubah sinyal suara menjadi teks atau perintah yang dapat dimengerti oleh mesin. Teknologi ini memainkan peran penting dalam interaksi antara manusia dan mesin, seperti pada sistem asisten virtual atau aplikasi biometrik untuk identifikasi pengguna (Dwijayanti dkk., 2022). Secara sederhana, *speech recognition* adalah cara untuk menerjemahkan suara manusia menjadi kata-kata dengan bantuan program komputer dan algoritma canggih. Menurut Alsobhani dkk. (2021), tujuan utama *speech recognition* adalah membuat mesin mampu mengenali suara manusia dan merespons sesuai dengan

perintah yang diberikan, sehingga menjadi langkah penting dalam menciptakan mesin yang cerdas dan mampu memahami manusia.

Pengenalan suara bisa dikategorikan berdasarkan jenis dan durasi sinyal suara yang dianalisis. Misalnya, ada pengenalan untuk kata-kata terisolasi, seperti "start" atau "stop," dan pengenalan untuk ucapan yang lebih panjang atau terhubung, seperti kalimat penuh. Teknologi ini menjadi semakin relevan dalam kehidupan sehari-hari karena membantu menjembatani komunikasi manusia-mesin dengan lebih natural melalui bahasa yang digunakan sehari-hari (Alsobhani dkk., 2021).

2.1.3 *Mel-frequency cepstral coefficients* (MFCC)

Mel-frequency cepstral coefficients (MFCC) adalah metode yang umum digunakan untuk mengubah suara menjadi representasi numerik yang dapat dipahami oleh komputer (Dwijayanti dkk., 2022). Ini adalah cara untuk merepresentasikan fitur dari suara dengan fokus pada aspek yang mirip dengan cara manusia mendengar. Menurut Musaev dkk. (2019), prosesnya dimulai dengan mengubah sinyal suara menjadi data yang menggambarkan frekuensi suara. Kemudian, data tersebut diterapkan pada skala mel, yang merupakan cara untuk mengukur frekuensi dengan cara yang lebih sesuai dengan sensitivitas telinga manusia terhadap suara.

MFCC berguna karena dapat menangkap berbagai elemen penting dalam suara, seperti intonasi (naik-turunnya nada) dan timbre (warna suara). Elemen-elemen ini sangat penting untuk membedakan satu suara dengan suara lainnya, seperti dalam pengenalan kata atau identifikasi pembicara (Talai dkk., 2023). Dalam dunia

pengenalan suara, MFCC adalah salah satu metode paling efektif untuk merepresentasikan ciri khas dari suara yang sedang dianalisis.

Karena kemampuannya menangkap informasi penting seperti timbre dan intonasi, MFCC banyak digunakan dalam berbagai aplikasi pengenalan suara, dari sistem asisten virtual hingga pengenalan suara dalam bahasa tertentu. Dalam banyak algoritma, terutama yang menggunakan jaringan saraf, MFCC menjadi dasar untuk menganalisis dan mengenali berbagai jenis suara. Dengan memanfaatkan fitur ini, model bisa lebih akurat dalam mengenali kata atau suara, bahkan dalam kondisi noise atau lingkungan yang tidak ideal.

Menurut Aini dkk. (2021), proses ekstraksi MFCC dimulai dengan pengambilan sampel dari sinyal suara yang ada. Langkah selanjutnya adalah *pre-emphasis*, yaitu memperkuat frekuensi tinggi agar suara lebih jelas dan lebih mudah dipisahkan dari noise. Kemudian, sinyal dibagi menjadi potongan-potongan kecil melalui proses *windowing* agar analisis bisa lebih detail. Setelah itu, *Fast Fourier Transform* (FFT) digunakan untuk mengubah sinyal dari domain waktu ke domain frekuensi. Di tahap berikutnya, *filter bank* yang disesuaikan dengan skala mel diterapkan untuk mengekstrak fitur-fitur frekuensi yang lebih sesuai dengan cara manusia mendengar. Terakhir, dilakukan skala logaritmik dan diterapkan *Discrete Cosine Transform* (DCT) untuk mengubah data menjadi koefisien cepstral yang lebih ringkas dan mewakili karakteristik suara secara efisien.

2.1.4 *Deep Learning*

Deep Learning merupakan cabang dari *machine learning* yang menggunakan arsitektur jaringan syaraf tiruan untuk memproses data. Metode ini terinspirasi dari cara neuron bekerja didalam otak manusia. Dalam *deep learning*, model dirancang menggunakan jaringan syaraf yang terdiri dari banyak lapisan yang dapat secara otomatis mengekstrak pola atau fitur dari data, bahkan yang tidak terstruktur melalui pelatihan berulang pada data tersebut (Gupta dkk., 2022).

Menurut Hernández-Blanco dkk. (2019), setiap lapisan dalam *deep learning* memiliki neuron yang memproses data dengan cara mengubahnya menggunakan perhitungan tertentu, baik yang bersifat linear maupun non-linier. *Deep learning* bekerja dengan cara melibatkan pembelajaran bertahap, dimana model dapat memahami informasi dari yang sederhana hingga yang lebih kompleks.

Salah satu teknik yang sering digunakan dalam *deep learning* adalah *Convolutional Neural Network* (CNN). *Deep learning* telah menunjukkan keunggulan besar dibandingkan metode machine learning tradisional, terutama dalam tugas-tugas seperti pengenalan pola, pengolahan gambar, dan aplikasi dalam bidang visi komputer. Keunggulan ini membuat *deep learning* menjadi teknologi kunci di balik berbagai inovasi, seperti pengenalan wajah, kendaraan otonom, dan asisten virtual (Gupta dkk., 2022).

2.1.5 *Convolutional Neural Network* (CNN)

Convolutional Neural Network (CNN) adalah jenis jaringan syaraf tiruan yang dirancang khusus untuk mengolah data dengan struktur grid, seperti gambar atau

suara (Khrisne & Hendrawati, 2020). CNN bekerja dengan cara yang mirip dengan cara kita melihat gambar, yaitu dengan fokus pada bagian-bagian kecil dari data untuk mengekstrak informasi penting. Jaringan ini memiliki beberapa lapisan utama, termasuk lapisan konvolusi, lapisan aktivasi (seperti ReLU), dan lapisan *pooling*, yang berfungsi untuk menangkap fitur penting dari input dengan cara yang efisien (Dwijayanti dkk., 2022). CNN (*Convolutional Neural Network*) memiliki keunggulan dalam mengklasifikasikan gambar dan suara yang kompleks dengan banyak karakter dan parameter. Kemampuan ini sangat berguna karena CNN dapat secara otomatis mengekstrak dan mengenali pola-pola penting dari data latih, sehingga memudahkan proses klasifikasi. Dengan arsitektur yang dirancang untuk mengenali berbagai fitur, CNN dapat mengidentifikasi perbedaan halus dan menghasilkan akurasi yang tinggi dalam klasifikasi data (Mawaddah dkk., 2021).

Pada tahap awal, CNN menggunakan lapisan konvolusi untuk menerapkan filter yang berfungsi mengekstrak berbagai fitur dari sinyal input. Fitur-fitur ini bisa berupa pola tertentu dalam gambar atau aspek suara yang lebih mendalam. Setelah proses konvolusi, ada lapisan *pooling* yang berfungsi untuk mengurangi dimensi data yang dihasilkan, sehingga membuat proses pelatihan menjadi lebih cepat dan model menjadi lebih efisien. Pengurangan dimensi ini juga membantu model untuk lebih mudah menggeneralisasi data baru, bukan hanya menghafal data yang sudah ada (Talai dkk., 2023).

Menurut Alsobhani dkk. (2021), CNN memiliki keunggulan utama yaitu kemampuannya dalam mengurangi dampak *noise* (gangguan), berbagi bobot di berbagai bagian input, serta mencegah model dari *overfitting* (memahami data

dengan terlalu mendalam). Dengan berbagi bobot, model menjadi lebih tahan terhadap variasi dalam data, baik itu dalam pengenalan gambar maupun suara. Dalam prosesnya, CNN melewati beberapa tahap berikut.

Arsitektur CNN memungkinkan model untuk belajar dengan cara yang sangat efisien dari data yang diberikan, menjadikannya sangat akurat dalam tugas-tugas klasifikasi suara atau pengenalan suara, bahkan ketika lingkungan penuh dengan noise atau variasi. Sebelum memahami cara kerja *convolutional neural networks* (CNN), penting untuk mengetahui bahwa sistem ini terdiri dari tiga jenis lapisan utama (Team Algoritma, 2022). Lapisan pertama, yang disebut *convolutional layer*, adalah lapisan awal yang bertugas mengenali elemen-elemen dasar seperti pola sederhana pada gambar atau suara.

Kemudian, ada *pooling layer*, yang juga dikenal sebagai downsampling layer. Lapisan ini melanjutkan proses dari lapisan pertama dengan menyederhanakan dan mempertegas hasil identifikasi untuk mendapatkan informasi yang lebih mendalam tentang objek visual atau audio. Terakhir, ada *fully-connected layer*, yang merupakan bagian paling kompleks dan inti dari CNN. Jika *convolutional layer* hanya bertugas mengenali elemen-elemen dasar seperti warna atau tekstur, *fully connected layer* bertugas untuk mengidentifikasi bentuk dan objek secara keseluruhan, sehingga mampu mengenali objek yang dimaksud dengan lebih tepat.

1. *Input Layer*

Input layer merupakan langkah awal untuk menerima data masukan, yaitu sinyal suara yang akan diolah oleh model CNN (Dua dkk., 2022). Lapisan ini

bertugas menyesuaikan dimensi data agar sesuai dengan struktur jaringan yang telah dirancang. Data yang diterima di input layer biasanya berupa matriks yang merepresentasikan fitur-fitur dari sinyal suara yang akan dianalisis (Alsobhani dkk., 2021). Proses ini sangat penting karena CNN membutuhkan input dalam format yang terorganisir agar dapat mengenali pola dengan lebih efektif dan akurat (Khrisne & Hendrawati, 2020).

2. *Convolutional Layer*

Convolutional layer adalah bagian utama dalam arsitektur *Convolutional Neural Network* (CNN) yang bertugas mengekstrak fitur dari data input, seperti gambar atau spektrogram. Proses ini melibatkan operasi konvolusi, di mana matriks filter (*feature detector*) digeser di atas data input untuk menghasilkan peta fitur (*feature map*). Peta fitur ini kemudian diproses lebih lanjut dengan lapisan aktivasi, seperti ReLU (*Rectified Linear Unit*), untuk meningkatkan nonlinieritas data (Dwijayanti dkk., 2022).

Menurut Khrisne & Hendrawati (2020), *convolutional layer* merupakan bagian penting dari arsitektur *Convolutional Neural Network* (CNN) yang digunakan untuk pengenalan suara. *Convolutional layer* berfungsi untuk mengekstrak fitur dari input suara yang telah diolah menggunakan *Mel Frequency Cepstral Coefficients* (MFCC). Proses ini melibatkan penerapan filter atau kernel pada input untuk menghasilkan peta fitur yang dapat membantu model dalam mengenali pola suara. Setiap *convolutional layer* dalam model CNN dapat memiliki beberapa filter yang berbeda, yang memungkinkan model untuk menangkap berbagai fitur dari data

suara. Setelah melalui beberapa *convolutional layer,* hasilnya akan diproses lebih lanjut melalui lapisan *pooling* dan *fully connected* untuk klasifikasi akhir.

3. *Activation Layer*

Menurut Azis dkk. (2021), *activation layer* adalah bagian dari jaringan saraf yang berfungsi menerapkan fungsi aktivasi pada output neuron. Fungsi ini bertujuan untuk menambahkan elemen non-linearitas ke dalam model, sehingga jaringan dapat mempelajari pola yang kompleks dan tidak linier dari data.

Activation layer biasanya digunakan setelah lapisan konvolusi atau *fully connected*. Dengan adanya fungsi aktivasi, model mampu mengenali pola yang lebih rumit dalam data, yang sangat penting untuk tugas klasifikasi. Penggunaan *activation layer* dengan fungsi yang sesuai membantu jaringan menangkap hubungan yang lebih mendalam dalam data, meningkatkan kemampuannya untuk memahami data baru, dan mempercepat proses pelatihan dengan membantu model mencapai konvergensi lebih cepat. Berikut beberapa fungsi aktivasi yang sering digunakan dalam CNN (Talai dkk., 2023).

a. *Rectified Linear Unit*

Atau disebut ReLU, fungsi ini mengubah nilai negatif menjadi nol, sementara nilai positif dibiarkan apa adanya. ReLU sering dipilih karena mampu mengatasi masalah vanishing gradient dan mempercepat proses konvergensi dalam pelatihan model.

b. Sigmoid

Fungsi ini mengubah output menjadi nilai di antara 0 dan 1. Biasanya digunakan pada lapisan output untuk klasifikasi biner, karena cocok untuk memprediksi probabilitas dua kelas.

c. Softmax

Fungsi ini digunakan di lapisan output untuk klasifikasi multi-kelas. Softmax mengubah output menjadi probabilitas untuk setiap kelas, di mana total probabilitas seluruh kelas berjumlah 1.

4. *Pooling Layer*

Pooling layer adalah salah satu komponen dalam arsitektur Convolutional Neural Network (CNN) yang berfungsi untuk mengecilkan dimensi peta fitur (*feature map*) yang dihasilkan oleh lapisan konvolusi. Proses ini membantu mengurangi jumlah parameter dan beban komputasi pada jaringan, sekaligus mengurangi risiko *overfitting* (Dua dkk., 2022).

Pooling layer memiliki beberapa parameter penting, seperti ukuran filter dan stride. Ukuran filter menentukan area yang akan diproses, sedangkan stride mengatur seberapa jauh filter bergerak melintasi *feature map*. Sebagai contoh, filter berukuran 3x3 dengan stride 2 akan bergeser sejauh 2 piksel setiap kali. Dengan adanya *pooling layer*, model CNN dapat mengurangi jumlah parameter yang perlu dipelajari, sehingga proses pelatihan menjadi lebih cepat dan model mampu menangani data baru dengan lebih baik (Azis dkk., 2021).

5. *Faltten Layer*

Flatten layer adalah proses yang mengubah data dari bentuk dua dimensi menjadi satu dimensi (vektor tunggal). Langkah ini dilakukan agar data dapat diteruskan ke lapisan berikutnya, khususnya *fully connected layer*. Fungsi utama *flatten layer* adalah merubah *feature map* yang dihasilkan oleh lapisan sebelumnya, seperti convolutional atau *pooling layers*, menjadi format satu dimensi (Azis dkk., 2021).

Proses ini sangat penting karena *fully connected layer* membutuhkan data dalam bentuk satu dimensi untuk melakukan klasifikasi. *Flatten layer* berperan sebagai penghubung antara lapisan konvolusi atau *pooling* dengan *fully connected layer*, yang merupakan tahap akhir dalam proses klasifikasi pada arsitektur CNN. Tanpa *flatten layer*, data tidak dapat diproses secara langsung oleh *fully connected layer*.

6. *Fully Connected Layer*

Menurut Talai dkk. (2023), *fully connected layer* adalah lapisan dalam jaringan saraf tiruan di mana setiap neuron terhubung dengan semua neuron pada lapisan sebelumnya. Artinya, setiap neuron menerima input dari semua neuron sebelumnya dan mengirimkan output ke lapisan berikutnya. Lapisan ini bertugas mengolah data yang telah diproses oleh lapisan-lapisan sebelumnya, seperti *convolutional* dan *pooling layers*, agar dapat digunakan untuk membuat keputusan akhir dalam klasifikasi. Lapisan ini mengubah data yang dihasilkan menjadi format yang cocok untuk menghasilkan output akhir.

Fully connected layer memungkinkan model untuk memahami pola yang kompleks dan menggabungkan informasi dari berbagai fitur. Namun, lapisan ini juga dapat meningkatkan jumlah parameter dalam model, yang berisiko menyebabkan *overfitting* jika tidak ditangani dengan baik.

2.1.6 Preprocessing

Menurut Shevira dkk. (2022), *preprocessing* adalah tahap di mana data dipersiapkan dan diolah agar sesuai dengan format yang diperlukan untuk diproses lebih lanjut. Tahapan ini dapat bervariasi tergantung pada kebutuhan, tetapi umumnya mencakup beberapa langkah seperti membersihkan data (misalnya, menghapus karakter yang tidak diperlukan), mengubah semua huruf menjadi kecil (*lowercase*), normalisasi, menghapus kata-kata umum yang kurang relevan (*stop-words*), mengubah kata ke bentuk dasarnya (*stemming*), dan memecah teks menjadi unit-unit kecil seperti kata atau frasa (*tokenizing*).

Tujuan dari proses ini adalah untuk menghilangkan elemen-elemen yang tidak penting atau mengganggu (*noise*), menyamakan format penulisan, serta membuat data lebih mudah dianalisis, misalnya dengan mengubah kata-kata slang atau ejaan yang tidak baku menjadi bentuk standar.

Menurut Roy dkk. (2018), *preprocessing* adalah langkah penting dalam mempersiapkan data sebelum dilakukan analisis lebih lanjut. Dalam bidang biologi sistem, *preprocessing* mencakup berbagai teknik untuk mengatasi data yang tidak lengkap, membersihkan noise, dan menangani nilai yang hilang. Proses ini melibatkan tahapan seperti pembersihan data, normalisasi, dan pemilihan fitur.

Semua langkah tersebut dilakukan untuk memastikan bahwa data sudah dalam kondisi optimal dan siap digunakan untuk analisis berikutnya.

Didalam *preprocessing,* proses ekstraksi fitur telah dilakukan untuk mendapatkan nilai MFCC *feature* nya. Setelah nilai fitur didapatkan, maka selanjutnya adalah melakukan normalisasi, *reshaping,* dan augmentasi pada data.

1. Normalisasi

Normalisasi data adalah langkah penting dalam pengolahan data, terutama untuk meningkatkan kualitas data yang digunakan dalam algoritma pembelajaran mesin. Proses ini bertujuan untuk menyelaraskan skala nilai pada data sehingga atribut-atribut dengan rentang nilai yang berbeda dapat dibandingkan dengan lebih mudah dan akurat (Permana & Salisah, 2022).

2. *Reshaping*

Reshaping adalah proses menyesuaikan dimensi atau struktur data, terutama dalam konteks citra atau data lainnya, agar sesuai dengan format yang diperlukan oleh model pembelajaran mesin atau deep learning. Langkah ini tidak mengubah isi utama dari data, melainkan hanya mengatur ulang bentuknya sehingga data dapat diproses dengan baik oleh model (Putri Ayuni dkk., 2023).

3. Augmentasi

Augmentasi adalah teknik yang digunakan untuk menambah jumlah data dalam dataset dengan memodifikasi data asli. Tujuan utama augmentasi adalah untuk memperkaya variasi dalam dataset, sehingga model dapat belajar lebih baik dan meningkatkan akurasi saat melakukan klasifikasi (Putri Ayuni dkk., 2023).

2.1.7 Training

Menurut Musaev dkk. (2019), *training* dalam konteks jurnal tersebut mengacu pada tahap penting dalam pengembangan model untuk pengenalan suara menggunakan jaringan saraf konvolusional (CNN). Proses ini melibatkan pembelajaran model dari data yang diberikan dengan tujuan mengenali pola-pola tertentu dan menghasilkan prediksi yang akurat.

Dalam jaringan saraf seperti CNN, *training* adalah proses inti di mana model dilatih untuk memahami pola dan karakteristik dari data input. Selama tahap ini, model mempelajari hubungan antara data input dan output yang diharapkan, sehingga mampu membuat prediksi atau klasifikasi berdasarkan data baru di masa depan. Proses ini sangat penting untuk memastikan model dapat berfungsi dengan baik dalam mengenali dan memahami data yang diberikan (Dua dkk., 2022).

2.1.8 Testing

Menurut Zulhaedi (2023), *testing* adalah kegiatan yang dilakukan untuk mengevaluasi parameter atau kemampuan suatu program atau sistem, sekaligus memastikan apakah hasilnya sesuai dengan kebutuhan atau harapan. Proses ini melibatkan pengujian untuk mengidentifikasi masalah dalam sistem sehingga dapat diperbaiki sebelum diluncurkan.

Melalui *testing*, kita dapat menemukan kekurangan pada perangkat lunak atau sistem yang perlu diperbaiki. *Testing* juga memastikan setiap komponen berfungsi dengan baik. Setelah komponen-komponen kecil digabungkan, dilakukan

pengujian menyeluruh pada hasil penggabungan tersebut, yang dikenal sebagai *Integration Testing*.

2.1.9 Python

Python adalah bahasa pemrograman yang bersifat interaktif dan dapat dijalankan di berbagai platform atau aplikasi. Python juga didukung oleh banyak library, yang sangat berguna untuk pengolahan data dan pengembangan machine learning (Mujilahwati dkk., 2021).

Menurut Budhi Gustiandi (2023), Python fleksibel dalam mendukung berbagai gaya pemrograman. Python menggunakan pendekatan *object-oriented programming* (OOP) untuk bekerja dengan kelas dan objek, atau menggunakan *functional programming* untuk menulis kode yang lebih modular dan berbasis fungsi. Fleksibilitas inilah yang membuat Python menjadi pilihan populer di kalangan pengembang dan data scientist.

2.1.10 Tensorflow

TensorFlow adalah platform open-source yang lengkap untuk *Machine Learning*. Platform ini memiliki ekosistem yang luas dan fleksibel, terdiri dari berbagai alat, pustaka, dan sumber daya komunitas yang membantu pengembangan teknologi *machine learning* terkini. Dengan TensorFlow, pengembang dapat dengan mudah membangun dan menyebarkan aplikasi *machine learning* (Anggeli dkk., 2021).

TensorFlow dirancang oleh tim Google Brain dan berfungsi sebagai pustaka matematika simbolis yang menggunakan konsep dataflow dan pemrograman

berbasis grafik untuk komputasi numerik. TensorFlow sangat cocok digunakan untuk proyek-proyek *machine learning* berskala besar. Platform ini memudahkan proses pengambilan data, pelatihan model, prediksi, serta penyempurnaan hasil yang diperoleh, sehingga sangat mendukung pengembangan model machine learning yang akurat dan efisien.

2.1.11 Confusion Matrix

Confusion matrix adalah alat yang digunakan untuk mengevaluasi kinerja model pengenalan suara yang telah dibangun (Khrisne & Hendrawati, 2020). Matriks ini memberikan gambaran tentang jumlah prediksi yang benar dan salah untuk setiap kelas, sehingga memudahkan peneliti dalam mengidentifikasi kesalahan yang dilakukan oleh model. Melalui confusion matrix, berbagai metrik evaluasi seperti akurasi, presisi, *recall*, dan F1-*score* dapat dihitung. Metrik-metrik ini memberikan pemahaman yang lebih mendalam tentang seberapa efektif sistem dalam mengenali sinyal suara, termasuk suara tonal maupun yang tidak biasa (Dua dkk., 2022).

1. Akurasi adalah rasio jumlah prediksi yang benar terhadap seluruh jumlah prediksi yang dilakukan oleh model (Tan, P.-N. dkk., 2006).
2. Presisi adalah rasio antara jumlah data positif yang diklasifikasikan dengan benar (*True Positive*) terhadap seluruh data yang diprediksi positif oleh model (*True Positive* + *False Positive*)(Han, J. dkk., 2011).
3. *Recall* mengukur seberapa banyak data positif yang berhasil ditemukan oleh model, dihitung dari rasio antara *True Positive* dan seluruh data aktual positif (*True Positive* + *False Negative*)(Chollet, F., 2018).

4. *F1-score* adalah rata-rata harmonik dari presisi dan *recall*, digunakan untuk mengevaluasi keseimbangan antara keduanya, terutama dalam kondisi data yang tidak seimbang (Géron, A., 2019).

2.2 Kajian Penelitian

Kajian Penelitian adalah bagian dari suatu karya ilmiah yang berfungsi untuk menampilkan dan menganalisis penelitian-penelitian sebelumnya yang relevan dengan topik yang sedang diteliti. Beberapa analisa penelitian terdahulu:

Tabel 2. 1 Kajian penelitian

No.	Judul	Peneliti	Identitas Jurnal
1.	Speech Recognition using Convolution Deep Neural Networks	Ayad Alsobhani, Hanaa M A Alabboodi, & Haider Mahdi	Journal of Physics: Conference Series 1973 (2021) 012166 doi:10.1088/1742-6596/1973/1/012166
2.	Implementasi Speech Recognition Pada Aplikasi E-Prescribing Menggunakan Algoritme Convolutional Neural Network	Nur Azis, Herwanto, & Fathurrahman Ramadhani	Jurnal Media Informatika Budidarma Volume 5, Nomor 2, April 2021, Page 460-467
3.	Structural Support Vector Machine for Speech Recognition Classification with CNN Approach	Kuldeep Chouhan, Abhishek Singh, Anurag Shrivastava, Shweta Agrawal, Brahma Datta Shukla, & Pragya Singh Tomar	Institute of Electrical and Electronics Engineers Inc. 2021 9th International Conference on Cyber and IT Service Management, CITSM 2021
4.	Developing a Speech Recognition System for Recognizing Tonal Speech Signals Using a Convolutional Neural Network	Sakshi Dua, Sethuraman Sambath Kumar, Yasser Albagory, Rajakumar Ramalingam, Ankur Dumka, Rajesh Singh, Mamoon Rashid, Anita Gehlot, Sultan S. Alshamrani, & Ahmed Saeed AlGhamdi	Appl. Sci. 2022, 12, 6223. https://doi.org/10.3390/app12126223 https://www.mdpi.com/journal/applsci
5.	Speaker Identification Using a	Suci Dwijayanti, Alvio Yunita Putri,	JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi) Vol. 6 No. 1 (2022)

No.	Judul	Peneliti	Identitas Jurnal
	Convolutional Neural Network	& Bhakti Yudho Suprapto	140 - 145 ISSN Media Electronic: 2580-0760
6.	Enhanced speech recognition for indonesian geographic dictionary using deep learning	H. Hugeng, D. Gunawan, & A. T. Kusumo	International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019
7.	Indonesian Alphabet Speech Recognition for Early Literacy using Convolutional Neural Network Approach (2020)	Duman Care Khrisne & Theresia Hendrawati,	Journal of Electrical, Electronics and Informatics, p-ISSN: 2549–8304 e-ISSN: 2622–0393 Vol. 4 No. 1, February 2020
8.	Image Approach to Speech Recognition on CNN (2019)	Muhammadjon Musaev, Ilyos Khujayorov, & Mannon Ochilov	ISCSIC 2019, September 25–27, 2019, ACM. ACM ISBN 978-1-4503-7661-7/19/09 https://doi.org/10.1145/3386164.3389100
9.	Pemodelan CNN Untuk Deteksi Emosi Berbasis Speech Bahasa Indonesia (2021)	Yulistia Khoirotul Aini, Tri Budi Santoso, &Titon Dutono	Jurnal Komputer Terapan Vol. 7, No. 1, Mei 2021, 143 – 152
10.	Comparative Study of CNN Structures for Arabic Speech Recognition (2023)	Zoubir Talai, Nada Kherici, & Halima Bahi	Ingénierie des Systèmes d'Information Vol. 28, No. 2, April, 2023, pp. 327-333 Journal homepage: http://iieta.org/journals/isi

Penelitian yang dilakukan oleh Ayad Alsobhani, Hanaa M A Alabboodi, & Haider Mahdi (2021) yang berjudul "*Speech Recognition using Convolution Deep Neural Networks*", penelitian ini menggunakan *Convolutional Neural Network* (CNN) untuk pengenalan suara, dengan data yang dikumpulkan di berbagai lokasi dengan tingkat kebisingan yang berbeda, termasuk pasar, rumah, taman, dan laboratorium. Penelitian ini menggunakan dua jenis kata: kata terisolasi (seperti "stop" dan "start") dan kata terhubung (seperti "backward"). Hasilnya arsitektur CNN-VGG-f dapat menghasilkan akurasi tinggi sebesar 98.78%, jauh lebih tinggi dibandingkan dengan metode lain seperti *Mel-frequency cepstral coefficients*

(MFCC) yang hanya mencapai 34.62% dan kombinasi MFCC dengan deltas yang mencapai 26.92%.

Penelitian yang dilakukan oleh Nur Azis, Herwanto, & Fathurrahman Ramadhani (2021) dengan judul "Implementasi *Speech Recognition* Pada Aplikasi E-Prescribing Menggunakan *Algoritme Convolutional Neural Network*", mengikuti tahapan terstruktur, dimulai dari identifikasi masalah hingga pengujian aplikasi, dengan diagram alur yang menggambarkan proses seperti pembuatan rancangan tampilan dan penerapan algoritma CNN. Ekstraksi ciri dilakukan menggunakan metode Mel-Frequency Cepstral Coefficients (MFCCs) dari file suara berformat .wav, yang kemudian diindeks ke dalam file JSON untuk proses pelatihan. Hasilnya dataset suara berjumlah 65.721 file dilatih menggunakan algoritme Convolutional Neural Network (CNN) dengan 40 epochs, menghasilkan akurasi sebesar 93% saat pelatihan dan 96% saat pengujian dalam kondisi nyata.

Penelitian yang dilakukan oleh Kuldeep Chouhan, Abhishek Singh, Anurag Shrivastava, Shweta Agrawal, Brahma Datta Shukla, & Pragya Singh Tomar (2021) yang berjudul "*Structural Support Vector Machine for Speech Recognition Classification with CNN Approach*", menggunakan kombinasi Structural Support Vector Machine (SSVM) dan Convolutional Neural Network (CNN) untuk meningkatkan akurasi dalam pengenalan suara. Tahapan penelitian meliputi pengumpulan data suara untuk melatih model, ekstraksi fitur dari sinyal suara, pelatihan model menggunakan SSVM, dan pengujian kinerja model menggunakan metrik seperti True Positive Rate (TP), True Negative Rate (TN), dan Equal Error Rate (EER). Hasilnya, metode SSVM dan CNN memberikan akurasi yang baik

dalam pengenalan suara, meskipun hasil terbaik dicapai oleh Random Forest (RF) dengan EER sebesar 0.56%, mengungguli SVM (1.93%) dan *Deep Neural Network* (DNN, 1.85%). Analisis menggunakan kurva ROC mengonfirmasi performa tinggi sistem dalam mendeteksi pengguna autentik dan tidak autentik.

Penelitian yang dilakukan oleh Sakshi Dua, Sethuraman Sambath Kumar, Yasser Albagory, Rajakumar Ramalingam, Ankur Dumka, Rajesh Singh, Mamoon Rashid, Anita Gehlot, Sultan S. Alshamrani, & Ahmed Saeed AlGhamdi (2022) yang berjudul "*Developing a Speech Recognition System for Recognizing Tonal Speech Signals Using a Convolutional Neural Network*", menggunakan metode Convolutional Neural Network (CNN) untuk mengembangkan sistem pengenalan suara menjadi teks yang mampu mengenali sinyal suara tonal dalam bahasa Punjabi. Dataset baru yang direkam melibatkan sinyal suara dari 11 pembicara dengan berbagai usia, aksen, dan lingkungan, termasuk suara latar tambahan seperti musik instrumental. Hasilnya, teknik ekstraksi fitur yang digunakan adalah Mel-Frequency Cepstral Coefficients (MFCC). Akurasi sebesar 89,15% dan tingkat kesalahan kata (Word Error Rate, WER) sebesar 10,56%, yang membuktikan bahwa penerapan CNN pada sinyal suara tonal menghasilkan kinerja yang lebih baik dibandingkan metode sebelumnya.

Penelitian yang dilakukan oleh Suci Dwijayanti, Alvio Yunita Putri, & Bhakti Yudho Suprapto (2022) yang berjudul "*Speaker Identification Using a Convolutional Neural Network*", mengembangkan metode identifikasi pembicara yang efektif menggunakan Convolutional Neural Network (CNN) dan spektrogram sebagai representasi fitur dari sinyal suara. Data yang digunakan berasal dari 78

pembicara, masing-masing mengucapkan tiga digit terakhir dari nomor identifikasi mahasiswa mereka dalam bahasa Indonesia sebanyak 10 kali, menghasilkan 780 data suara. Hasilnya, model CNN yang digunakan mencapai akurasi klasifikasi sebesar 97.06%, dan semakin banyak data yang diberikan, semakin baik akurasi klasifikasi yang dihasilkan oleh CNN.

Penelitian yang dilakukan oleh H. Hugeng, D. Gunawan, & A. T. Kusumo (2019) yang berjudul "*Enhanced speech recognition for indonesian geographic dictionary using deep learning*", menggunakan Automatic Speech Recognition (ASR) berbasis Convolutional Neural Network (CNN) untuk meningkatkan akurasi pengenalan suara pada kamus geografi Indonesia. Dataset terdiri dari 50 kata geografi yang diucapkan oleh 20 mahasiswa (19 laki-laki dan 1 perempuan), menghasilkan total 5.000 data suara. Model CNN dibangun menggunakan Python dan TensorFlow, dengan uji coba dilakukan pada dua jenis input: kata terisolasi dan kata kontinu. Hasilnya, CNN mencapai akurasi rata-rata 80% untuk kata terisolasi dan 72,67% untuk kata kontinu, jauh lebih tinggi dibandingkan metode GMM-HMM, yang hanya mencapai akurasi 52,87% pada kata terisolasi. Pada kata terisolasi, akurasi tertinggi adalah 96%, sedangkan untuk kata kontinu mencapai 84%.

Penelitian yang dilakukan oleh Duman Care Khrisne & Theresia Hendrawati (2020) yang berjudul "*Indonesian Alphabet Speech Recognition for Early Literacy using Convolutional Neural Network Approach*", menggunakan pendekatan pengenalan suara dengan metode Convolutional Neural Network (CNN), di mana data input berupa suara yang diolah menggunakan fitur vektor

suara Mel Frequency Cepstral Coefficients (MFCC) untuk membentuk matriks 3-dimensi yang digunakan sebagai input ke dalam CNN dengan arsitektur model Sequential yang terdiri dari 10 lapisan sederhana. Hasilnya, model CNN mampu mengenali suara dengan akurasi tinggi sebesar 84% dan F-Measure sebesar 0.91, dengan ukuran model yang cukup kecil, sekitar 6 MB.

Penelitian yang dilakukan oleh Muhammadjon Musaev, Ilyos Khujayorov, & Mannon Ochilov (2019) yang berjudul "*Image Approach to Speech Recognition on CNN*", menggunakan pendekatan pengenalan suara berbasis gambar dengan memanfaatkan jaringan saraf konvolusional (CNN), di mana karakteristik spektral dari sinyal suara diubah menjadi gambar menggunakan algoritma Short-Time Fourier Transform (STFT). Gambar-gambar ini kemudian dianalisis untuk mengklasifikasikan suara berdasarkan fitur-fitur penting yang dipilih. Hasilnya, sistem pengenalan fonem berbasis CNN dapat mencapai akurasi antara 72.4% hingga 77.5% pada dataset TIMIT, dan penggunaan teknik pengolahan citra, seperti Scale-invariant Feature Transform (SIFT), dapat meningkatkan klasifikasi suara yang terisolasi.

Penelitian yang dilakukan oleh Yulistia Khoirotul Aini, Tri Budi Santoso, &Titon Dutono (2021) yang berjudul "Pemodelan CNN Untuk Deteksi Emosi Berbasis *Speech* Bahasa Indonesia", mengembangkan sistem Speech Emotion Recognition (SER) berbasis bahasa Indonesia, menggunakan dataset dari TV series "Imperfect." Proses dimulai dari pengumpulan, pre-processing, dan pelabelan data audio menjadi kategori emosi (marah, senang, netral, sedih) dengan frekuensi sampling 44100 Hz. Fitur suara yang diekstraksi meliputi Mel Frequency Cepstral

Coefficients (MFCC), frekuensi fundamental, dan Root Mean Square Energy (RMSE). Hasilnya, Kombinasi MFCC dan frekuensi fundamental memberikan akurasi tertinggi sebesar 85%, sedangkan kombinasi MFCC dan RMSE memiliki akurasi terendah, yaitu 72%, karena RMSE tidak merepresentasikan emosi dengan baik.

Penelitian yang dilakukan oleh Zoubir Talai, Nada Kherici, & Halima Bahi (2023) yang berjudul "*Comparative Study of CNN Structures for Arabic Speech Recognition*", membandingkan berbagai arsitektur Convolutional Neural Network (CNN) untuk pengenalan suara bahasa Arab, guna menentukan model yang paling cocok serta memberikan panduan bagi pengembangan sistem pengenalan suara di masa depan, khususnya untuk bahasa dengan sumber daya terbatas. Hasilnya, GoogLeNet mencapai akurasi tertinggi sebesar 89.61%, diikuti oleh AlexNet dengan 86.19%, dan ResNet dengan 83.46%, yang membuktikan bahwa CNN dapat secara efektif memodelkan fitur akustik untuk bahasa Arab.

Dari berbagai penelitian yang menggunakan *Convolutional Neural Network* (CNN) untuk pengenalan suara, dapat disimpulkan bahwa CNN adalah metode yang sangat unggul dibandingkan pendekatan tradisional seperti MFCC dan GMM-HMM. CNN mampu memahami fitur kompleks dari suara, sehingga sangat efektif untuk berbagai aplikasi, seperti pengenalan suara umum, identifikasi pembicara, hingga pengenalan emosi. Tingkat akurasi yang dicapai cukup tinggi, antara 72% hingga 98%, tergantung pada desain model dan data yang digunakan. Misalnya, penelitian oleh Duman Care Khrisne & Theresia Hendrawati berhasil mengenali

suara dengan akurasi tinggi sebesar 84%, sementara Nur Azis dkk. memperoleh 96% untuk pengujian aplikasi berbasis suara.

Penelitian-penelitian tersebut juga menunjukkan bahwa pemilihan fitur suara sangat berpengaruh pada kinerja model. Fitur seperti MFCC sering memberikan hasil yang terbaik, sementara tambahan seperti frekuensi fundamental dapat meningkatkan akurasi lebih jauh. Secara keseluruhan, CNN membuka peluang besar dalam pengembangan teknologi berbasis suara, khususnya untuk bahasa dan aplikasi yang sebelumnya dianggap sulit diimplementasikan, seperti pendidikan, teknologi asisten pintar, dan pengenalan suara di lingkungan berisik.

Berdasarkan hasil dari penelitian-penelitian sebelumnya, kebanyakan studi pengenalan suara masih berfokus pada penggunaan alfabet bahasa Latin atau bahasa lain yang lebih umum. Penelitian tentang pengenalan suara huruf Hiragana masih jarang dilakukan. Penelitian ini berusaha menonjolkan perbedaannya dengan mengembangkan sistem pengenalan suara khusus untuk huruf Hiragana, sehingga memperluas cakupan pengenalan suara ke bahasa Jepang, yang memiliki pola pengucapan dan struktur bunyi yang unik.

Selain itu, beberapa penelitian sebelumnya belum banyak membahas pentingnya pengurangan *noise* dalam proses pengenalan suara. Untuk mengatasi masalah ini, penelitian ini menggunakan teknik pengurangan *noise* pada tahap *preprocessing* guna membersihkan suara dari gangguan. Dengan data suara yang lebih bersih, model dapat mempelajari pola suara dengan lebih baik dan meningkatkan akurasi pengenalan huruf Hiragana.