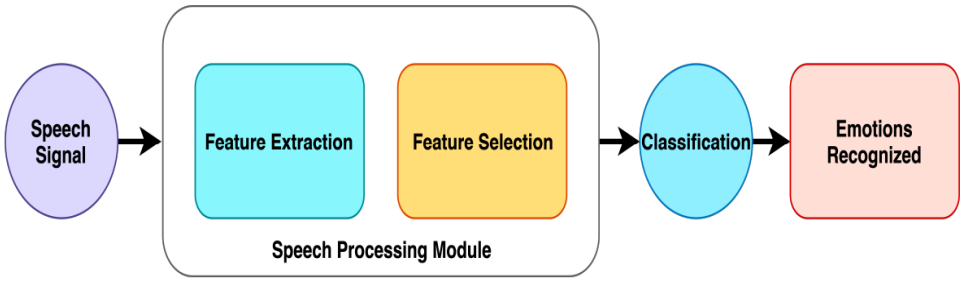
BAB II

TINJAUAN PUSTAKA

2.1 Landasan Teori

2.1.1 Speech Emotion Recognition (SER)

Speech Emotion Recognition (SER) dapat didefinisikan sebagai suatu sistem pengenalan pola. Hal ini menunjukkan bahwa tahapan-tahapan yang terdapat pada sistem pengenalan pola juga terdapat pada sistem pengenalan emosi berbasis ucapan (Selvaraj dkk., 2016). Gambar 2.1 merupakan alur dari sistem yang disederhanakan yang digunakan untuk *Speech Emotion Recognition.*



Gambar 2.1 Tahapan SER

Berdasarkan Gambar 2.1, *Speech Emotion Recognition* berisi lima modul utama yaitu, input ucapan emosional, ekstraksi fitur, pemilihan fitur, klasifikasi, dan output emosional yang dikenali (El Ayadi dkk., 2011).

Proses untuk mengenali sekumpulan emosi secara otomatis menjadi perhatian utama dalam SER. Oleh karena itu untuk mengklasifikasikan sejumlah besar data emosi sangatlah rumit. Masalah SER telah dibahas selama beberapa tahun menggunakan metode statistik dan algoritma *machine*

learning, seperti *Support Vector Machines* (SVMs) dan berbagai algoritma regresi (Tarantino dkk., 2019).

2.1.2 Klasifikasi

Klasifikasi adalah proses menemukan suatu model yang menggambarkan dan membedakan kelas data atau konsep, dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas dari objek yang label kelasnya belum diketahui. Selain itu, klasifikasi digunakan secara luas pada *data mining* untuk mengelompokkan data ke dalam beragam kelas (Hasanain & Rizal, 2021). Teknik dari klasifikasi adalah dengan melihat variabel dari kelompok data yang sudah ada. Klasifikasi bertujuan untuk memprediksi kelas dari suatu objek yang tidak diketahui sebelumnya. Umumnya klasifikasi terdiri dari tiga tahap, yaitu pembangunan model, penerapan model, dan evaluasi (Nasution dkk., 2019).

2.1.3 Convolutional Neural Network (CNN)

CNN adalah pengembangan dari *Multilayer Perceptron* (MLP) yang didesain untuk mengolah data dua dimensi. CNN termasuk dalam jenis *Deep Neural Network* karena kedalaman jaringan yang tinggi dan banyak diaplikasikan pada data data dua dimensi (Nurona Cahya dkk., 2021). Cara kerja CNN memiliki kesamaan pada MLP, namun dalam CNN setiap neuron dipresentasikan dalam bentuk dua dimensi, tidak seperti MLP yang setiap neuron hanya berukuran satu dimensi (Hasanain & Rizal, 2021).

Menurut (Martiyaningsih, 2022) secara umum CNN memiliki beberapa layer, yaitu *Convolutional Layer, Pooling Layer*, dan *Fully Connected Layer.*

1. *Convolutional Layer*, merupakan tahap dimana seluruh data menyentuh lapisan *convolutional* yang mengalami proses konvolusi dan akan difilter kemudian menghasilkan sebuah *activation map*. Lapisan pada layer ini memiliki tiga parameter, yaitu *depth, stride, dan zero padding*.
2. *Pooling Layer,* merupakan lapisan yang menggunakan *Feature Map* sebagai input dan mengolahnya berdasar nilai piksel terdekat. Lapisan ini disisipkan ke dalam lapisan konvolusi secara teratur. Bentuk Pooling Layer yang paling umum adalah 2x2.
3. *Fully Connected Layer,* merupakan lapisan yang digunakan untuk tujuan melakukan transformasi pada dimensi data agar dapat diklasifikasikan secara linear.

2.1.4 Mel Frequency Cepstral Coefficient (MFCC)

Mel Frequency Cepstral Coefficient (MFCC) merupakan metode ekstraksi fitur untuk mendapatkan lebih banyak informasi dalam data audio. MFCC dapat digunakan untuk mengekstrak data sinyal suara sehingga bisa didapatkan ciri yang terdapat pada data sinyal suara (Emanuella dkk., 2021). MFCC bekerja menyerupai adaptasi sistem pendengaran manusia, dengan cara memfilter sinyal suara secara logaritmik untuk frekuensi tinggi yang nilainya di atas 1000 Hz dan memfilter sinyal suara secara linear untuk frekuensi rendah yang nilainya di bawah 1000 Hz (Danika dkk., 2023). Metode ini efektif dalam menangkap karakteristik penting dari suara karena didasarkan pada persepsi pendengaran manusia.

Proses ekstraksi fitur MFCC terdiri dari beberapa tahapan penting. Dimulai dengan *pre-emphasis* untuk meningkatkan energi sinyal pada frekuensi tinggi, dilanjutkan dengan *framing* yang membagi sinyal menjadi frame-frame pendek. Selanjutnya, *windowing* dilakukan untuk mengurangi diskontinuitas di tepi *frame*. *Fast Fourier Transform* (FFT) kemudian digunakan untuk mengubah sinyal dari domain waktu ke domain frekuensi (Gupta and Gupta, 2016; Ralev and Krastev, 2023) .

Setelah transformasi ke domain frekuensi, spektrum diproses melalui serangkaian filter segitiga yang terdistribusi secara mel, yang dikenal sebagai *Mel Filter Bank*. Langkah berikutnya adalah *logarithmic compression* untuk menyesuaikan dengan persepsi manusia terhadap loudness. Kemudian, *Discrete Cosine Transform* (DCT) digunakan untuk mengubah log mel spectrum menjadi *domain cepstral* (Sabitha V, 2013; Gupta and Gupta, 2016).

MFCC telah terbukti efektif dalam berbagai aplikasi pemrosesan suara. Dalam konteks pengenalan emosi, MFCC mampu menangkap perubahan dalam nada suara dan intonasi yang sering dikaitkan dengan ekspresi emosional. Penelitian menunjukkan bahwa MFCC, ketika dikombinasikan dengan algoritma pembelajaran mesin seperti *Support Vector Machine* (SVM) atau *Convolutional Neural Network* (CNN), dapat mencapai akurasi yang tinggi dalam klasifikasi emosi (Ravi and Taran, 2024). Meskipun MFCC sangat efektif, beberapa penelitian telah mengusulkan modifikasi atau alternatif untuk meningkatkan kinerjanya. Misalnya, *Power-Normalized Cepstral*

Coefficients (PNCC) dan *Modified Group Delay Function* (ModGDF) telah diusulkan sebagai alternatif yang potensial, terutama dalam kondisi berderau.

Dalam implementasinya, MFCC telah diterapkan dalam berbagai domain di luar pemrosesan suara manusia. Termasuk klasifikasi suara lingkungan, deteksi kebocoran dalam jaringan distribusi air, dan bahkan dalam analisis getaran untuk diagnosis kesalahan mekanis. Keunggulan MFCC terletak pada kemampuannya untuk menangkap karakteristik spektral suara yang relevan dengan persepsi manusia (Razak et al., 2008; das, Jena and Barik, 2014).

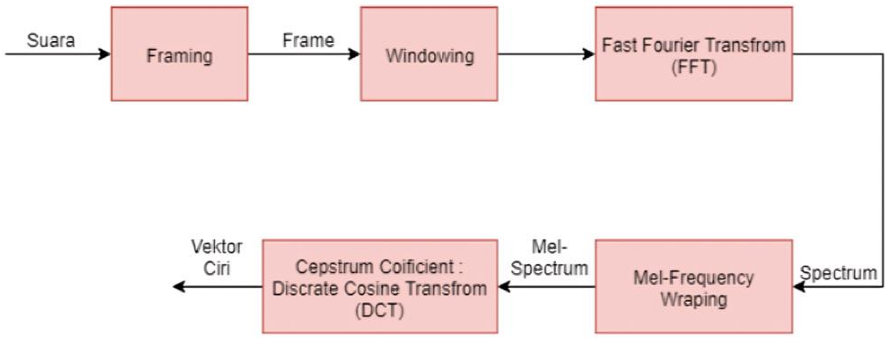
Dengan perkembangan teknologi pembelajaran mendalam, MFCC sering digunakan sebagai input untuk arsitektur *neural network* yang kompleks. Misalnya, kombinasi MFCC dengan *Long Short-Term Memory* (LSTM) *networks* telah menunjukkan hasil yang menjanjikan dalam tugas-tugas seperti pengenalan emosi. Selain itu, penggunaan MFCC bersama dengan teknik augmentasi data seperti *gaussian noise, time stretching*, dan *pitch shifting* telah terbukti dapat meningkatkan kinerja model (Padman and Magare, 2023; Neili and Sundaraj, 2024).

Dalam konteks pengenalan emosi dari suara, MFCC telah menunjukkan keefektifannya ketika digabungkan dengan berbagai teknik pembelajaran mesin. Penelitian terbaru menunjukkan bahwa penggunaan MFCC bersama dengan model seperti BiLSTM atau CNN dapat mencapai akurasi yang tinggi dalam klasifikasi emosi. Beberapa studi bahkan menggabungkan MFCC dengan fitur tekstual dan visual untuk meningkatkan akurasi klasifikasi emosi

dalam konteks multimodal (Mishra, Warule and Deb, 2024; Nikhila Gudipati et al., 2024).

MFCC juga telah diaplikasikan dalam berbagai bidang lain seperti klasifikasi *genre* musik, deteksi suara sintetis, dan bahkan dalam pemantauan kesehatan. Dalam klasifikasi genre musik, MFCC telah digunakan untuk mengekstrak fitur dari lagu-lagu dalam berbagai bahasa, termasuk Bengali. Sementara itu, dalam konteks kesehatan, MFCC telah digunakan untuk menganalisis suara batuk dan pola pernapasan untuk diagnosis penyakit pernapasan (Khan and Hafiz, 2024; Sai Varun et al., 2024).

Meskipun MFCC telah terbukti sangat efektif, penelitian terus berlanjut untuk meningkatkan kinerjanya dan mengeksplorasi kombinasinya dengan teknik-teknik baru. Beberapa penelitian terbaru fokus pada optimalisasi parameter MFCC, penggabungannya dengan fitur lain seperti spektrogram, dan penggunaan teknik pembelajaran mesin yang lebih canggih untuk meningkatkan akurasi klasifikasi. Dengan perkembangan ini, MFCC terus menjadi salah satu metode ekstraksi fitur yang paling penting dan banyak digunakan dalam pemrosesan sinyal suara (Zhong, 2023; Bai, 2024; Gourisaria et al., 2024).



Gambar 2.2 Alur Ekstraksi Fitur MFCC

Gambar 2.2 menjelaskan alur dari tahapan ekstraksi fitur MFCC. Pertama, sinyal suara dipotong untuk menghilangkan keheningan atau gangguan yang mungkin muncul pada awal maupun akhir suara dilakukan *framing*, membagi ke dalam sejumlah frame. Selanjutnya, proses *windowing* digunakan untuk meminimalkan diskontinuitas sinyal. *Fast Fourier Transform* (FFT) kemudian diterapkan untuk mengubah setiap frame ke domain frekuensi. Dalam *Mel-frequency wrapping block*, sinyal diplot terhadap spektrum Mel untuk meniru pendengaran manusia. Terakhir, *Discrete Cosine Transform* (DCT) dilakukan untuk menghasilkan vektor ciri (Handoko & Suyanto, 2019).

2.1.5 *Crowd-sourced Emotional Multimodal Actors Dataset* (CREMA-D)

Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) adalah *dataset* berlabel yang digunakan untuk mempelajari pengenalan ekspresi dan emosi. CREMA-D mencakup 7.442 klip dari 91 aktor dan aktris dengan usia dan etnis yang beragam dengan mengekspresikan enam bentuk emosi secara universal yaitu *anger, disgust, fear, happy, neutral and sad* (Cao dkk., 2014).

Secara umum, usia dan latar belakang etnis dapat memengaruhi cara seseorang berbicara, termasuk pola naik-turun nada suaranya (intonasi). Menurut Jacewicz dan Fox (2013), perbedaan usia memengaruhi anatomi dan fungsi pita suara, misalnya kelenturan dan ketebalan lipatan vokal, sehingga nada dasar serta rentang frekuensi suara bisa berbeda pada tiap kelompok umur. Di sisi lain, latar belakang etnis berkaitan dengan dialek atau aksen tertentu, yang juga membentuk ciri khas intonasi setiap kelompok budaya (Laver, 1994). Dengan demikian, kombinasi antara faktor usia dan etnis dapat menghasilkan karakter suara yang beragam, termasuk pola intonasi yang berbeda.

Dalam SER, latar belakang etnis dan dialek berhubungan erat dengan variasi cara berbicara, seperti perbedaan aksen, pilihan kata, dan pola naik-turun nada (intonasi). Faktor-faktor ini dapat memengaruhi ciri akustik yang diandalkan oleh sistem SER, misalnya frekuensi dasar suara (pitch) atau kecepatan berbicara, sehingga model yang tidak dilatih dengan data dari berbagai etnis dan dialek berisiko kesulitan mengenali emosi pada penutur tertentu. Dalam salah satu penelitian ditunjukkan bahwa perbedaan aksen dapat membuat pola suara marah atau senang terdengar berbeda, sehingga model SER yang hanya berfokus pada satu dialek atau aksen tertentu cenderung menghasilkan akurasi yang lebih rendah bila diterapkan pada penutur dengan latar belakang bahasa atau etnis lain (Elbarougy et al., 2014). Dengan demikian, variasi etnis dan dialek penting diperhatikan untuk memastikan bahwa sistem SER mampu mengakomodasi seluruh pengguna secara andal.

2.2 Penelitian Terkait dan Kebaruan Penelitian

2.2.1 State of The Art

Tabel 2.1 menunjukan perbandingan penelitian yang berhubungan dengan fokus pada permasalan penelitian dan hasil penelitian yang dihasilkan dalam proses klasifikasi.

Tabel 2.1 State of The Art

No	Nama Peneliti	Tahun	Judul	Masalah	Hasil
1	Rizqi Fathin Fadhillah, Raden Sumiharto	2023	Klasifikasi Suara Untuk Memonitori Hutan Berbasis *Convolutional Neural Network*	Beberapa upaya untuk mencegah aktivitas ilegal di hutan telah dilakukan seperti patroli oleh petugas setempat. Akan tetapi upaya tesebut mengalami beberapa	Berdasarkan hasil penelitian ini, didapatkan bahwa model MobileNetV3-Small + MLP dengan data latih gabungan dari augmentasi time stretch dan time shift memberikan performa yang paling bagus

No	Nama Peneliti	Tahun	Judul	Masalah	Hasil
				kendala, diantaranya adalah sumber daya manusia yang kurang, sarana dan prasarana yang kurang mendukung, dan keterbatasan dana. Patroli yang dilakukan oleh manusia juga terbatas pada waktu tertentu, sehingga pada waktu yang lain masih berpotensi adanya tindakan ilegal.	untuk diimplementasikan pada sistem ini, dengan durasi inference 0.8 detk; akurasi sebesar 93.96%; dan presisi sebesar 94.1%.

No	Nama Peneliti	Tahun	Judul	Masalah	Hasil
2	Andani Achmad, Adnan, Muhammad Rijal	2022	Klasifikasi Penyakit Pernapasan Berbasis Visualisasi Suara Menggunakan Metode *Support Vector Machine*	Saat ini, sistem kesehatan pernafasan telah memanfaatkan metode machine learning untuk melakukan klasifikasi penyakit pernapasan melalui visualisasi suara stetoskop yang merupakan informasi grafik bunyi dalam rentang waktu dan frekuensi tertentu.	Hasil pengujian mengemukakan bahwa metode Support Vector Machine berhasil diterapkan dalam mengklasifikasi penyakit asma, bronkitis dan tuberkulosis dengan akurasi sebesar 46.37%.

No	Nama Peneliti	Tahun	Judul	Masalah	Hasil
3	Cecilia Tania Emanuella, Musfita, dan Armin Lawi	2021	Klasifikasi Suara Kucing dan Anjing Menggunakan Convolutional Neural Network	Kucing dan anjing merupakan salah satu hewan peliharaan manusia yang berkomunikasi melalui suara. Untuk membantu mengetahui, membedakan dan mengenali suara hewan dengan peliharaan yang lain, dibutuhkan klasifikasi suara hewan agar lebih jelas.	Metode praproses dan metode klasifikasi dengan menggunakan Convolutional Neural Network cukup baik untuk menentukan kebenaran dari klasifikasi data audio. Hal ini terbukti dengan hasil akurasi sebesar 88%. Perubahan tingkat confusion tidak mempengaruhi hasil akurasi. Hal ini

No	Nama Peneliti	Tahun	Judul	Masalah	Hasil
					membuktikan bahwa klasifikasi menggunakan metode CNN relatif baik terhadap perubahan parameter yang dilakukan.
4	ABHISHEK SEHGAL AND NASSER KEHTARNAVAZ	2021	A Convolutional Neural Network Smartphone App for Real-Time Voice Activity Detection	VAD digunakan untuk tujuan di mana klasifikasi atau estimasi kebisingan diaktifkan oleh VAD untuk menyesuaikan parameter algoritma pengurangan derau tergantung pada kelas	Makalah ini telah menyediakan jaringan saraf tiruan convolutional untuk melakukan deteksi aktivitas suara secara realtime dengan latensi audio yang rendah. Aplikasi ini telah

No	Nama Peneliti	Tahun	Judul	Masalah	Hasil
				atau jenis derau. Untuk bagian sinyal atau bagian di mana ucapan dalam derau atau ucapan+derau terdeteksi, tidak ada klasifikasi/estimasi derau yang dilakukan dan pengurangan derau dilakukan berdasarkan jenis derau yang diidentifikasi terakhir.	dikembangkan untuk smartphone Android dan iOS. Arsitektur dari jaringan saraf convolutional telah dioptimalkan untuk memungkinkan bingkai audio untuk diproses secara real-time tanpa bingkai apa pun yang dilewati dengan tetap mempertahankan akurasi suara yang tinggi deteksi aktivitas.

No	Nama Peneliti	Tahun	Judul	Masalah	Hasil
5	Muhammad Hasbi Ashshiddieqy, Jondri, Achmad Rizal	2020	Klasifikasi Suara Paru Dengan Convolutional Neural Network (CNN)	Deteksi kelainan pada suara paru biasa dilakukan dengan menggunakan stetoskop. Pendeteksian suara paru dengan stetoskop memiliki keterbatasan karena sangat bergantung pada individu yang melakukan pemeriksaan.	Setelah melakukan pelatihan pada model pembelajaran mesin CNN dengan menggunakan generalisasi berupa augmentasi dan dropout layer, maka didapatkan akurasi sebesar 84,80% terhadap data latih dan 78,09% terhadap data uji.
6	Mazin Abed Mohammed, Karrar	2020	Voice Pathology Detection and Classification Using	Patologi suara memiliki dampak negatif pada	Hasil eksperimen menunjukkan metode CNN

No	Nama Peneliti	Tahun	Judul	Masalah	Hasil
	Hameed Abdulkareem, Salama A. Mostafa, Mohd Khanapi Abd Ghani, Mashael S. Maashi, Begonya Garcia-Zapirain, Ibon Oleagordia, Hosam Alhakami and Fahad Taha AL-Dhief		Convolutional Neural Network Model	keteraturan getaran dan fungsi suara, yang mengarah pada peningkatan kebisingan suara. Suara normal berubah menjadi tegang, lemah dan serak yang mempengaruhi kualitas. Sampai saat ini, metode deteksi patologi suara yang ada saat ini memiliki evaluasi yang bias	yang diusulkan untuk deteksi patologi ucapan mencapai akurasi hingga 95,41%. Metode ini juga memperoleh 94,22% dan 96,13% untuk F1-Score dan Recall. Sistem yang diusulkan menunjukkan kemampuan yang tinggi dari aplikasi klinis nyata yang menawarkan solusi diagnosis dan perawatan

No	Nama Peneliti	Tahun	Judul	Masalah	Hasil
				berdasarkan hal-hal yang bersifat subjektif.	otomatis yang cepat dalam waktu 3 detik untuk mencapai akurasi klasifikasi.
7	Andre Danika Jangkung Raharjo Bambang Hidayat	2020	SENTIMENT ANALYSIS ONLINE SHOP ON THE PLAY STORE USING METHOD SUPPORT VECTOR MACHINE (SVM)	Perbedaan utama kedua jenis senar adalah suara yang dihasilkan. Senar baja cenderung menghasilkan suara yang lebih nyaring, sedangkan senar nylon cenderung menghasilkan suara yang mellow. Selain	Hasil dari proses tersebut kemudian akan diklasifikasikan dengan metode support vector machine (SVM) dengan fungsi kernel RBF sebagai fungsi terbaik dengan akurasi 95%. Gitar senar

No	Nama Peneliti	Tahun	Judul	Masalah	Hasil
				itu senar baja juga menghasilkan volume suara yang lebih besar dibandingkan dengan suara yang dihasilkan senar nylon.	baja cenderung menghasilkan frekuensi maksimum yang lebih besar dibandingkan dengan senar nylon.
8	Raditya Budi Handoko dan Suyanto	2019	Klasifikasi Gender Berdasarkan Suara Menggunakan Support Vector Machine	Banyak metode telah diusulkan untuk membangun sistem klasifikasi gender berbasis suara yang berakurasi tinggi, namun umumnya masih	Sistem klasifikasi gender berdasarkan suara menggunakan SVM yang dibangun mampu memberikan akurasi sebesar 100%. Pemilihan parameter dan

No	Nama Peneliti	Tahun	Judul	Masalah	Hasil
				kurang tahan terhadap derau atau noise.	kernel SVM sangat mempengaruhi akurasi sistem. Kernel RBF dan Sigmoid mempunyai akurasi lebih buruk dibanding Linear dan Polynomial (d =1).
9	Aditya Singgi Prayogi, Maulana Rizqi, Tresna Maulana Fahrudin	2019	Klasifikasi Suara Tangisan Bayi Berdasarkan Prosodic Features Menggunakan Metode Moments of	Bagi orang tua yang mengasuh bayi tentu sulit untuk menentukan apa yang diinginkan oleh bayi, karena suara tangisan yang	Akurasi terbaik pada proses klasifikasi menggunakan data sampling Percentage Rate yaitu 76% dimana nilai K yang digunakan adalah 9.

No	Nama Peneliti	Tahun	Judul	Masalah	Hasil
			Distribution dan K-Nearest Neighbours	dikeluarkan hampir sama. Padahal ada perbedaan arti tertentu dari suara tangisan bayi tersebut.	Sedangkan akurasi terbaik pada proses klasifikasi menggunakan data sampling Leave One Out yaitu 42% dengan nilai K yang digunakan adalah 5.
10	Nur Hudha Wijaya, Indah Soesanti, Eka Firmansyah	2017	KLASIFIKASI SUARA JANTUNG MENGGUNAKAN NEURAL NETWORK BACKPROPAGATION	Untuk melakukan diagnose suara jantuk normal atau abnormal (disebut murmur patologis) diperlukan kepekaan dan pengalaman oleh dokter, dengan	Pendekatan ciri statistis dengan menghitunng nilai mean, mode, variance, deviation, skewness, kurtosis, entropy klasifikasi dengan neural

No	Nama Peneliti	Tahun	Judul	Masalah	Hasil
			BERBASIS CIRI STATISTIS	demikian hasil diagnose sangat dipengaruhi oleh subjektivitas dokter.	backpropagation memberikan hasil Accuracy = 91,72%, Sensitivity = 99,50%, Spesificity = 79,17%, Precision = 90,16%. Berdasarkan hasil klasifikasi dengan metode artificial neural network backpropagation menunjukkan accuracy mencapai 91,72%.

Penelitian-penelitian sebelumnya dalam bidang pengenalan emosi berbasis suara telah menggunakan berbagai metode klasifikasi, seperti *Hidden Markov Model* (HMM), *Gaussian Mixture Model* (GMM), *Support Vector Machine* (SVM), dan metode *deep learning*, terutama *Convolutional Neural Network* (CNN) (Achmad dkk., 2022; Ashshiddieqy dkk., 2020; Danika dkk., 2023; Emanuella dkk., 2021; Fadhillah & Sumiharto, 2023; Handoko & Suyanto, 2019; Mohammed dkk., 2020; Prayogi dkk., 2019; Sehgal & Kehtarnavaz, 2018; Wijaya dkk., 2017). Meskipun model-model tradisional seperti HMM dan SVM telah digunakan dengan cukup baik dalam beberapa aplikasi, akurasi yang dihasilkan masih terbatas, khususnya ketika menghadapi dataset yang tidak seimbang dan heterogen. Sebaliknya, penggunaan CNN dalam penelitian lebih mutakhir menunjukkan hasil yang lebih menjanjikan, seperti pada penelitian menggunakan dataset SAVEE yang mencapai akurasi 88% pada data latih. Namun, model-model ini masih menghadapi masalah *overfitting* ketika diterapkan pada data uji, dengan akurasi yang menurun hingga 52%, yang menunjukkan kebutuhan akan pendekatan yang lebih robust.

Selain itu, banyak penelitian yang menggunakan teknik augmentasi data seperti *gaussian noise*, *time stretching*, dan *pitch shifting* untuk memperkaya variasi data latih dan meningkatkan kinerja model. Teknik-teknik ini telah terbukti efektif dalam mengatasi masalah ketidakseimbangan kelas emosi dalam dataset, serta meningkatkan kemampuan model untuk mengenali emosi dalam berbagai kondisi suara. Meskipun demikian, beberapa penelitian menunjukkan bahwa pengenalan emosi masih terbatas pada jenis-jenis emosi

yang paling dominan dalam dataset, sementara emosi yang lebih kompleks atau langka sering kali kurang terdeteksi. Oleh karena itu, masih ada tantangan besar dalam meningkatkan generalisasi model agar dapat mengenali berbagai emosi dengan lebih akurat pada data uji yang beragam.

Penelitian ini bertujuan untuk mengisi kesenjangan yang ada dengan menggunakan dataset CREMA-D, yang menawarkan variasi emosi dan intensitas suara yang lebih luas. Dataset ini lebih representatif karena mencakup berbagai variasi nada suara, dan situasi yang relevan untuk pengenalan emosi berbasis suara. Selain itu, penelitian ini mengusulkan penggunaan teknik augmentasi data yang lebih beragam untuk memperkaya data pelatihan, serta peningkatan metode ekstraksi fitur untuk mengatasi masalah overfitting dan ketidakseimbangan kelas dalam dataset. Dengan pendekatan ini, diharapkan penelitian ini dapat meningkatkan kinerja model CNN dalam mengenali emosi pada suara dengan lebih akurat, serta memberikan kontribusi signifikan untuk pengembangan teknologi *Speech Emotion Recognition (*SER) yang lebih efektif dan dapat digeneralisasi ke berbagai aplikasi nyata.

2.2.2 Matriks Penelitian

Matriks Penelitian merupakan matriks yang berisi gambaran keseluruhan isi penelitian terkait dan penelitian yang akan dilakukan. Matriks penelitian pada penelitian ini dapat dilihat pada Tabel 2.2.

Tabel 2.2 Matriks Penelitian

No	Penulis / Tahun	Judul	Metode					Tujuan		
			ANN	CNN	SVM	KNN	MFCC	Klasifikasi	Deteksi	Aplikasi
1	Rizqi Fathin Fadhillah, Raden Sumiharto (2023)	Klasifikasi Suara Untuk Memonitori Hutan Berbasis Convolutional Neural Network		√				√		
2	Andani Achmad, Adnan, Muhammad Rijal (2022)	KLASIFIKASI PENYAKIT PERNAPASAN BERBASIS VISUALISASI SUARA			√		√	√		

No	Penulis / Tahun	Judul	Metode					Tujuan		
			ANN	CNN	SVM	KNN	MFCC	Klasifikasi	Deteksi	Aplikasi
		MENGGUNAKAN METODE SUPPORT VECTOR MACHINE								
3	Andre Danika Jangkung Raharjo Bambang Hidayat (2022)	Deteksi Suara Gitar Dengan Bahan Jenis Senar Berbeda Melalui Ciri Akustik Dengan Mel-Frequency Cepstral Coefficients (MFCC) Dan Support Vector Machine (SVM)			√		√		√	

No	Penulis / Tahun	Judul	Metode					Tujuan		
			ANN	CNN	SVM	KNN	MFCC	Klasifikasi	Deteksi	Aplikasi
4	Cecilia Tania Emanuella, Musfita, dan Armin Lawi (2021)	Klasifikasi Suara Kucing dan Anjing Menggunakan Convolutional Neural Network		√				√		
5	ABHISHEK SEHGAL AND NASSER KEHTARNAV AZ 2021	A Convolutional Neural Network Smartphone App for Real-Time Voice Activity Detection		√						√
6	Muhammad Hasbi	Klasifikasi Suara Paru Dengan		√				√		

No	Penulis / Tahun	Judul	Metode					Tujuan		
			ANN	CNN	SVM	KNN	MFCC	Klasifikasi	Deteksi	Aplikasi
	Ashshiddieqy, Jondri, Achmad Rizal (2020)	Convolutional Neural Network (CNN)								
7	Mazin Abed Mohammed, Karrar Hameed Abdulkareem, Salama A. Mostafa, Mohd Khanapi Abd Ghani, Mashael S. Maashi,	Voice Pathology Detection and Classification Using Convolutional Neural Network Model		√				√	√	

No	Penulis / Tahun	Judul	Metode					Tujuan		
			ANN	CNN	SVM	KNN	MFCC	Klasifikasi	Deteksi	Aplikasi
	Begonya Garcia-Zapirain, Ibon Oleagordia, Hosam Alhakami and Fahad Taha AL-Dhief (2020)									
8	Raditya Budi Handoko dan Suyanto (2019)	Klasifikasi Gender Berdasarkan Suara Menggunakan Support Vector Machine			√		√	√		

No	Penulis / Tahun	Judul	Metode					Tujuan		
			ANN	CNN	SVM	KNN	MFCC	Klasifikasi	Deteksi	Aplikasi
9	Aditya Singgi Prayogi, Maulana Rizqi, Tresna Maulana Fahrudin (2019)	Klasifikasi Suara Tangisan Bayi Berdasarkan Prosodic Features Menggunakan Metode Moments of Distribution dan K-Nearest Neighbours				√		√		
10	Nur Hudha Wijaya, Indah Soesanti, Eka	KLASIFIKASI SUARA JANTUNG MENGGUNAKAN NEURAL	√`					√		

No	Penulis / Tahun	Judul	Metode					Tujuan		
			ANN	CNN	SVM	KNN	MFCC	Klasifikasi	Deteksi	Aplikasi
	Firmansyah (2017)	NETWORK BACKPROPAGATION BERBASIS CIRI STATISTIS								
	Dede Septa Maulana Fajar (2023)	KLASIFIKASI EMOSI BERDASARKAN SUARA PADA DATASET CREMA-D MENGGUNAKAN CONVOLUTIONAL		√`			√`	√		

No	Penulis / Tahun	Judul	Metode					Tujuan		
			ANN	CNN	SVM	KNN	MFCC	Klasifikasi	Deteksi	Aplikasi
		NEURAL NETWORK								

2.2.3 Relevansi Penelitian

Relevansi Penelitian merupakan keterkaitan antar variabel pada penelitian terkait dengan penelitian yang dilakukan. Relevansi Penelitian dapat dilihat pada Tabel 2.3.

Tabel 2.3 Relevansi Penelitian

Peneliti	(Khoirotul Aini, Budi Santoso and Dutono, 2021)	(Dede Septa Maulana Fajar, 2023)
Judul	Pemodelan CNN Untuk Deteksi Emosi Berbasis Speech Bahasa Indonesia	Klasifikasi Emosi Berdasarkan Suara pada Dataset Crema-D menggunakan *Convolutional Neural Network*

Peneliti	**(Khoirotul Aini, Budi Santoso and Dutono, 2021)**	**(Dede Septa Maulana Fajar, 2023)**
Masalah Penelitian	Sampai saat ini SER masih menghadapi beberapa tantangan seperti tingkat akurasi yang rendah dari pengklasifikasi yang digunakan, kompleksitas komputasi yang tinggi, dan kelangkaan dalam ketersediaan natural data sets. Pengenalan emosi ucapan adalah tugas yang sulit karena beberapa alasan seperti definisi emosi yang ambigu dan pemisahan yang tidak jelas antara emosi yang berbeda.	Penelitian dalam model klasifikasi dalam bidang SER telah banyak dilakukan, tetapi hanya sedikit diantara banyak penelitian bidang SER yang menghasilkan pekerjaan yang efektif dalam memprediksi adanya emosi melalui ucapan.
Objek Penelitian	Emosi berdasarkan suara manusia	Emosi berdasarkan suara manusia
Algoritma/Metode	*Convolutional Neural Network*	*Convolutional Neural Network* dan MFCC
Dataset	Data set pada penelitian ini diambil dari TV series berbahasa Indonesia berjudul “Imperfect”. Hal ini	Dataset yang digunakan dari CREMA-D yang mencakup 7.442 klip dari 91 aktor dan aktris

Peneliti	(Khoirotul Aini, Budi Santoso and Dutono, 2021)	(Dede Septa Maulana Fajar, 2023)
	dilakukan dengan pertimbangan bahwa data dari TV series memiliki pembagian dialog terstruktur dan kualitas audio yang baik.	dengan usia dan etnis yang beragam dengan mengekspresikan enam bentuk emosi secara universal yaitu *anger, disgust, fear, happy, neutral and sad.*

Berdasarkan Tabel 2.3 terdapat satu penelitian yang memiliki keterkaitan dengan penelitian yang dilakukan. Penelitian yang dilakukan memiliki latar belakang masalah, objek penelitian dan metode yang digunakan cukup memiliki keterkaitan, namun dari penelitian sebelumnya yang dilakukan perlu adanya penambahan jumlah dataset yang digunakan sehingga lebih banyak data yang diklasifikasikan. Oleh karena itu, penelitian ini menggunakan dataset dari CREMA-D yang merupakan suatu standarisasi dataset untuk melakukan klasifikasi emosi berdasarkan suara.