BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian terdahulu telah mengidentifikasi dan menggunakan metode dan klasifikasi yang beragam, dimana hal tersebut bisa dapat menjadi acuan untuk melakukan analisis terhadap metode metode tersebut. Berbagai macam metode yang digunakan oleh penelitian terdahulu yaitu KNN, SVM, Random Forest, Decision Tree, dan CNN.

2.1.1 A Novel Decision Tree for Depression Recognition in Speech [9]

Penelitian dengan judul "*A Novel Decision Tree for Depression Recognition in Speech*" mencakup berbagai penelitian sebelumnya yang menggunakan fitur berbasis ucapan atau kalimat untuk mengidentifikasi depresi. Proses ini menggunakan dataset dari MODMA dimana penelitian ini menggunakan dataset yang sama dengan yang akan diuji. Metode-metode yang telah diteliti meliputi penggunaan fitur produksi ucapan dan prosodi, pendekatan berbasis neural network, serta analisis prosodi yang terkait dengan depresi. Penelitian lain mencakup algoritme baru untuk ekstraksi fitur ucapan, penggunaan fitur *Mel-Frequency Cepstral Coefficients* (MFCC), dan penerapan algoritma *Self-Organizing Map* (SOM) untuk klasterisasi. Terdapat juga penelitian yang mengkaji efek dari fitur ucapan tertentu dalam mengidentifikasi depresi pada remaja. Beberapa studi telah mencoba meningkatkan kinerja model dengan menggunakan berbagai metode fusion, termasuk penggunaan metode deteksi depresi berbasis banyak bahasa, gabungan fitur manual dan deep learning, serta fusion model yang menggabungkan sinyal audio, video, dan teks. Penelitian ini menunjukkan bahwa indikator perilaku memiliki kontribusi yang berbeda terhadap deteksi depresi pada pria dan wanita. Dengan menguji secara terpisah, model dapat lebih akurat dalam mengidentifikasi depresi dengan mempertimbangkan perbedaan gender ini.

Metode yang diusulkan dalam dokumen ini berfokus pada pencarian metode fusion segmen ucapan yang memiliki kemampuan generalisasi yang kuat dan kinerja tinggi untuk identifikasi depresi, serta membandingkannya dengan metode fusion segmen ucapan yang telah diteliti sebelumnya.

2.1.2 Automated Depression Analysis Using Convolutional Neural Networks from Speech [8]

Penelitian yang telah dilakukan dalam upaya pengenalan depresi dari rekaman audio telah menggunakan berbagai metode, termasuk penggunaan fitur audio dasar yang diekstraksi dengan *toolkit open-source Emotion and Affect Recognition (*openEAR*)*, dan pendekatan *Support Vector Regression* (SVR). Penelitian lainnya telah mengadopsi kombinasi dari spektra *eigenvalue* dan fitur koordinasi untuk menganalisis hubungan antara perilaku vokal dan skala depresi, dengan penggunaan *Gaussian staircase regression* untuk prediksi skor BDI–II. Selain itu, juga telah dilakukan penelitian yang menggali hubungan antara prosodi vokal dan perubahan keparahan depresi dari waktu ke waktu, menunjukkan bahwa prosodi vokal merupakan alat yang berharga untuk analisis depresi.

Untuk AVEC2016 dan AVEC2017, organisator menyediakan file audio dan transkrip tetapi tidak menyediakan klip video asli. Fitur audio yang diekstraksi dengan toolkit COVAREP (v1.3.2) telah digunakan untuk prediksi depresi akhir.

2.1.3 Comparison of Pre-Trained CNNs for Audio Classification Using Transfer Learning [15]

Penelitian ini membandingkan beberapa CNN yang sudah dilatih sebelumnya untuk klasifikasi audio menggunakan transfer learning. Penelitian ini mengevaluasi tiga CNN yang dilatih untuk klasifikasi gambar (GoogLeNet, SqueezeNet, dan ShuffleNet) dan dua CNN yang

dilatih khusus untuk klasifikasi suara (VGGish dan YAMNet). CNN dilatih menggunakan tiga dataset suara publik: UrbanSound8K, ESC-10, dan Air Compressor. Penelitian ini mengeksplorasi pengaruh parameter pelatihan utama, termasuk optimizer, ukuran mini-batch, learning rate, dan jumlah epoch, terhadap akurasi klasifikasi dan waktu pemrosesan. Lalu untuk dataset yang digunakan serupa dan metode yang di gunakan yaitu ekstraksi dari audio ke gambar, yang dimana hal ini serupa dengan apa yang digunakan oleh peneliti. InceptionV3 merupakan arsitektur CNN yang dirancang dengan modularitas dan efisiensi komputasi yang tinggi. Arsitektur ini menggunakan beberapa inovasi untuk mengurangi beban komputasi dan meningkatkan kinerja, termasuk penggunaan konvolusi terfaktorisasi dan penggabungan keluaran dari berbagai ukuran filter. InceptionV3 adalah salah satu arsitektur Convolutional Neural Network (CNN) yang dikembangkan oleh Google. Arsitektur ini merupakan pengembangan dari versi sebelumnya, yaitu GoogLeNet atau InceptionV1 dan InceptionV2. InceptionV3 dirancang untuk meningkatkan efisiensi komputasi dan akurasi model dengan memperkenalkan berbagai inovasi teknis.

2.1.4 MODMA Dataset: A Multi-modal Open Dataset for Mental-disorder Analysis [7]

MODMA dataset adalah dataset terbuka multi-modal untuk analisis gangguan mental yang mencakup data EEG dan audio dari pasien depresi klinis dan kontrol normal yang cocok. Dataset ini mencakup sinyal EEG yang dikumpulkan menggunakan 128-elektroda cap elastis tradisional dan pengumpul EEG 3-elektroda baru untuk aplikasi pervasif. Sinyal EEG dari 53 subjek direkam dalam keadaan istirahat dan di bawah stimulasi; sinyal EEG 3-elektroda dari 55 subjek direkam dalam keadaan istirahat; data audio dari 52 subjek direkam selama wawancara, membaca, dan deskripsi gambar. Penelitian ini bertujuan untuk menyediakan data berkualitas tinggi untuk analisis gangguan mental, yang sulit didapatkan karena data fisiologis yang baik untuk pasien gangguan mental sulit diperoleh. Semua pasien dipilih dengan hati-hati oleh psikiater

profesional di rumah sakit dan mengikuti kriteria diagnostik yang ketat. Dataset ini diharapkan dapat digunakan oleh peneliti lain di bidang ini untuk menguji metode analisis gangguan mental mereka.

2.1.5 ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices [11]

Penelitian ini membahas arsitektur CNN yang sangat efisien bernama ShuffleNet, yang dirancang khusus untuk perangkat mobile dengan daya komputasi yang sangat terbatas. ShuffleNet menggunakan dua operasi baru, yaitu pointwise group convolution dan channel shuffle, untuk mengurangi biaya komputasi sambil mempertahankan akurasi. Pointwise group convolution mengurangi kompleksitas komputasi dari konvolusi 1x1, sedangkan channel shuffle memungkinkan aliran informasi antar fitur channel dengan efisien. Penelitian ini menunjukkan bahwa ShuffleNet menggunakan lebih sedikit komputasi dibandingkan model lain seperti MobileNet dan menunjukkan kecepatan yang jauh lebih tinggi dibandingkan AlexNet dengan akurasi yang sebanding.

2.1.6 Rethinking the Inception Architecture for Computer Vision [12]

Penelitian ini membahas arsitektur jaringan saraf konvolusional (CNN) baru yang dinamai InceptionV3, yang dirancang untuk meningkatkan efisiensi penggunaan sumber daya komputasi dalam jaringan. Arsitektur ini memungkinkan peningkatan kedalaman dan lebar jaringan tanpa meningkatkan anggaran komputasi secara signifikan. Arsitektur InceptionV3 menggunakan konvolusi 1x1, 3x3, dan 5x5, serta pengurangan dimensi melalui konvolusi untuk mengurangi kompleksitas komputasi. Teknik pengurangan dimensi ini membantu menjaga anggaran komputasi tetap rendah sambil meningkatkan kedalaman dan lebar jaringan. Selain itu, penelitian ini menunjukkan penggunaan klasifikasi tambahan di tahap tengah jaringan untuk meningkatkan sinyal gradien dan stabilitas pelatihan. Klasifikasi tambahan ini bertindak sebagai regularizer yang membantu meningkatkan konvergensi jaringan dan

mencegah overfitting. Arsitektur InceptionV3 dirancang untuk mudah diadaptasi dan dituning untuk set label lain, sehingga cocok untuk transfer learning. Penelitian ini menunjukkan bahwa InceptionV3 mampu mencapai hasil yang sangat baik dalam kompetisi ImageNet dengan efisiensi komputasi yang lebih baik dibandingkan model sebelumnya

Berdasarkan hasil penelitian terdahulu yang sudah di susun dan di teiliti oleh penulis, maka dibuatkan poin penting untuk menjadikan acuan penelitian untuk pengujian analisis ini.

Tabel 1.1 Tabel Data

Judul Penelitian (Tahun)	**Poin**
A Novel Decision Tree for Depression Recognition in Speech (2020)	● Mengusulkan metode decision tree baru untuk pengenalan depresi dari ucapan. ● Menekankan pendekatan berbasis gender yang efektif. ● Menggunakan dataset MODMA ● Train per pasien, bukan per label *Healthy Control* dan *Major depressive disorder*
Automated Depression Analysis Using Convolutional Neural Networks from Speech (2018)	● Menerapkan CNN untuk analisis otomatis depresi dari data audio ● Menggunakan open-source Emotion and Affect Recognition (openEAR) untuk ekstraksi fitur. ● Menyelidiki hubungan antara perilaku vokal dan skala depresi melalui fitur prosodi. ● Menggunakan Spektrum untuk proses klasifikasi CNN.
Comparison of Pre-Trained CNNs for Audio Classification Using Transfer Learning (2021)	● Membandingkan beberapa CNN yang sudah dilatih sebelumnya untuk klasifikasi audio menggunakan

	transfer learning. ● Optimizer Adam, SGDM. ● Epoch 6, 8, 10. ● ShuffleNet, GoogleNet, SqueezeNet, VGGish, Yamnet. ● menguji gambar dan suara. ● menggunakan dataset ImageNet, dan YouTube. ● Evaluasi tiga CNN yang dilatih untuk klasifikasi gambar dan dua CNN yang dilatih khusus untuk klasifikasi suara. ● Menggunakan tiga dataset suara publik: UrbanSound8K, ESC-10, dan Air Compressor. ● Menggunakan dataset serupa ● Mengeksplorasi pengaruh parameter pelatihan utama terhadap akurasi klasifikasi dan waktu pemrosesan.
MODMA Dataset: A Multi-modal Open Dataset for Mental-disorder Analysis (2019)	● Dataset multi-modal untuk analisis gangguan mental. ● Data audio dikumpulkan selama wawancara, membaca, dan deskripsi gambar. ● Dirancang untuk menyediakan data berkualitas tinggi untuk analisis gangguan mental.
ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices (2017)	● Mengembangkan arsitektur CNN yang sangat efisien untuk perangkat mobile. ● Menggunakan pointwise group convolution dan channel shuffle untuk mengurangi biaya komputasi. ● ShuffleNet menggunakan lebih sedikit komputasi dibandingkan model lain.

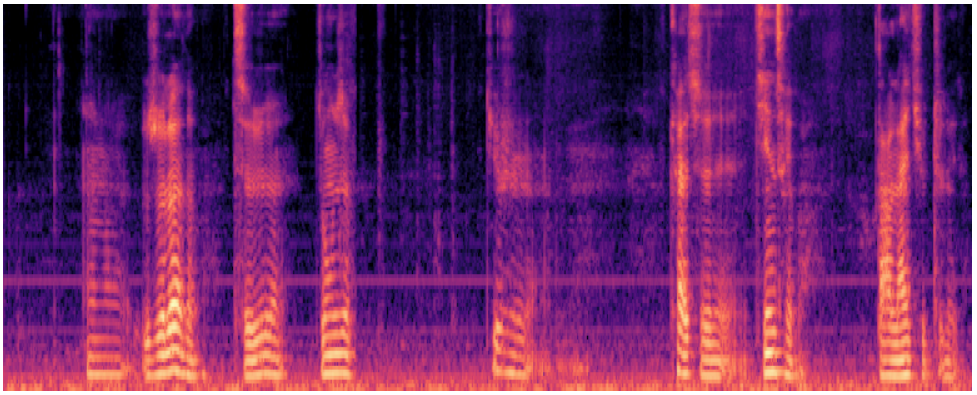
Rethinking the Inception Architecture for Computer Vision (2015)	● Mengusulkan arsitektur InceptionV3 untuk meningkatkan efisiensi komputasi. ● Menggunakan konvolusi terfaktorisasi untuk mengurangi beban komputasi. ● Menerapkan batch normalization untuk meningkatkan stabilitas pelatihan. ● Menggunakan auxiliary classifiers untuk memperbaiki konvergensi jaringan. ● Memperkenalkan teknik label smoothing untuk regularisasi model. ● Mencapai kinerja state-of-the-art dalam kompetisi ImageNet dengan efisiensi tinggi. ● Arsitektur ini menunjukkan bagaimana desain modular dapat meningkatkan kedalaman dan lebar jaringan tanpa peningkatan komputasi yang berlebihan. ● Menggunakan dataset ImageNet untuk pelatihan dan evaluasi. ● Dataset terdiri dari gambar, bukan audio. ● Fokus pada ekstraksi fitur dari gambar untuk meningkatkan akurasi klasifikasi. ●

2.2 Tinjauan Teori

2.2.1 Mel-Spectogram

Mel-spectrogram [8] dijelaskan sebagai metode visualisasi yang menggambarkan intensitas sinyal, atau tingkat kekerasan suara, dari suara yang direkam selama periode waktu tertentu, dengan memperlihatkan

variasi frekuensi dalam bentuk gelombang. Alat ini memungkinkan pengamatan perbandingan energi antara frekuensi yang berbeda, serta fluktuasi energi seiring berjalannya waktu. Pemilihan mel-spectogram sebagai masukan untuk Convolutional Neural Networks (CNN) dalam analisis sinyal audio didasarkan pada beberapa pertimbangan strategis. Mel-Spectogram mengubah sinyal audio linear menjadi format dua dimensi yang menggambarkan distribusi waktu dan frekuensi, dengan intensitas warna yang mencerminkan amplitudo. Transformasi ini menghasilkan representasi visual yang mendetail dan informatif, memungkinkan CNN untuk mengidentifikasi pola dan karakteristik suara dengan lebih efisien daripada sinyal audio mentah. Alasan ini menjadikan Mel-Spectogram pilihan yang lebih baik untuk aplikasi yang memerlukan pemahaman mendalam tentang konten audio, seperti pengenalan ucapan, analisis musik, atau deteksi emosi dari suara. Berikut adalah contoh bentuk mel-spectogram dari audio Gambar 2.1 mel-spectogram.



Gambar 2.1 Mel-Spectogram Train (MDD)

2.2.2 Deep Learning

Deep learning[13], yang merupakan sub dari *machine learning*, menggunakan jaringan saraf tiruan dengan banyak lapisan untuk meniru fungsi otak manusia dalam mengolah dan belajar dari data dalam jumlah besar. Teknologi ini memungkinkan model untuk secara otomatis dan secara efisien mengekstraksi fitur penting dari data mentah, tanpa perlu

pemrograman manual fitur oleh manusia. Dalam konteks Deteksi Depresi Berbasis Sinyal Audio, penerapan deep learning, khususnya melalui penggunaan Convolutional Neural Networks (CNN), menjadi sangat relevan.

2.2.3 Convolutional Neural Network (CNN)

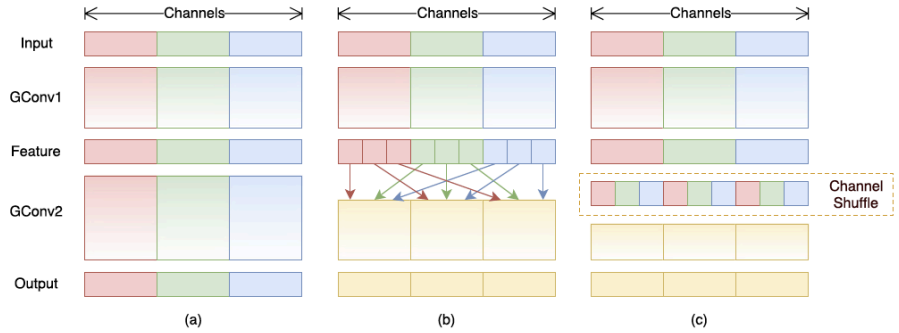
CNN adalah jenis jaringan saraf khusus yang sangat efektif dalam mengolah data yang memiliki struktur grid, seperti gambar atau, dalam kasus ini, Mel-Spectogram dari sinyal audio. Mel-Spectogram, yang mengonversi sinyal audio menjadi representasi visual dari frekuensi suara sepanjang waktu, menyediakan format dua dimensi yang kaya informasi. CNN dapat mengidentifikasi pola kompleks dalam Mel-Spectogram ini, seperti karakteristik suara yang mungkin menunjukkan depresi, dengan mempelajari fitur dari data tanpa perlu definisi fitur spesifik secara manual.

Menggunakan Mel-Spectogram sebagai input untuk CNN dalam deteksi depresi memanfaatkan kekuatan deep learning dalam mengenali dan mempelajari pola suara yang kompleks dan halus, yang mungkin tidak terdeteksi oleh metode tradisional. Hal ini memungkinkan untuk analisis yang lebih akurat dan mendalam dari sinyal audio, yang dapat meningkatkan kemampuan sistem dalam mengidentifikasi tanda-tanda depresi dengan lebih efektif. Kombinasi antara deep learning, CNN, dan analisis Mel-Spectogram membuka peluang baru dalam pengembangan solusi canggih untuk deteksi dini dan diagnosis kondisi kesehatan mental seperti depresi, dengan pendekatan yang lebih otomatis, objektif, dan tidak invasif.

2.2.4 ShuffleNet

ShuffleNet [11] merupakan arsitektur Convolutional Neural Network (CNN) yang dirancang khusus untuk meningkatkan efisiensi komputasi dengan mengurangi jumlah parameter tanpa mengorbankan akurasi. Arsitektur ini mengintegrasikan teknik pointwise group convolution dan

channel shuffle, yang secara signifikan memperkecil beban komputasi sambil mempertahankan kinerja model.



Gambar 2.2 ShuffleNet

Pointwise Group Convolution: Dalam ShuffleNet, pointwise group convolution digunakan untuk mengurangi kompleksitas komputasi dari konvolusi 1x1 yang biasanya mahal dalam hal sumber daya komputasi. Teknik ini membagi channel input menjadi grup dan melakukan konvolusi secara terpisah pada setiap grup.

Channel Shuffle: Untuk mengatasi keterbatasan dari group convolution, dimana grup-grup channel tidak saling berinteraksi, ShuffleNet memperkenalkan operasi channel shuffle. Operasi ini efektif mengatur ulang channel-channel sehingga grup yang berbeda dapat berbagi informasi, memungkinkan model untuk mempertahankan kemampuan representasi yang kuat.

ShuffleNet memanfaatkan kedua teknik ini untuk menciptakan jaringan yang sangat ringan namun efektif, cocok untuk aplikasi pada perangkat dengan sumber daya terbatas seperti ponsel pintar dan perangkat IoT. Model ini telah terbukti efektif pada dataset standar seperti ImageNet dan MS COCO, memberikan pendekatan yang sangat baik dalam menangani tugas-tugas pengenalan visual dengan lebih cepat dan efisien.

2.2.5 InceptionV3

InceptionV3[12] adalah arsitektur CNN inovatif yang memperkenalkan modul-modul Inception untuk meningkatkan kedalaman dan lebar jaringan tanpa memperbesar beban komputasi secara signifikan. Arsitektur ini memanfaatkan berbagai ukuran filter konvolusi (1x1, 3x3, dan 5x5) dan teknik reduksi dimensi untuk mengurangi kompleksitas komputasi. Dengan pendekatan ini, jaringan dapat lebih efisien dalam mengenali pola yang kompleks tanpa menambah jumlah parameter yang berlebihan.

Modul InceptionV3 menggabungkan filter konvolusi dari berbagai ukuran dalam satu blok, memungkinkan jaringan menangkap informasi pada berbagai skala secara bersamaan. Ini meningkatkan kemampuan jaringan dalam mengekstrak dan menggabungkan fitur penting dari berbagai ukuran area.

Efisiensi penggunaan sumber daya juga ditingkatkan melalui konvolusi 1x1 untuk reduksi dimensi sebelum konvolusi yang lebih besar, menjaga anggaran komputasi tetap rendah sambil meningkatkan kedalaman dan lebar jaringan.

Penelitian ini juga menunjukkan penggunaan klasifikasi tambahan di tahap tengah jaringan untuk meningkatkan sinyal gradien dan stabilitas pelatihan. Klasifikasi tambahan ini bertindak sebagai regularizer yang membantu meningkatkan konvergensi jaringan dan mencegah overfitting.

InceptionV3 telah menunjukkan hasil yang sangat baik dalam kompetisi ImageNet, menunjukkan peningkatan akurasi yang signifikan dibandingkan model sebelumnya serta efisiensi dalam penggunaan komputasi. Dengan lebih dari dua puluh lapisan, InceptionV3 mengilustrasikan bagaimana arsitektur yang dalam dan kompleks bisa

dijalankan dengan efisien bahkan pada perangkat dengan sumber daya terbatas.

Selain itu, arsitektur ini mengadopsi teknik regularisasi seperti label smoothing, yang membantu model untuk lebih tahan terhadap overfitting dengan mendistribusikan probabilitas ke seluruh label. InceptionV3 juga menggunakan batch normalization untuk meningkatkan kecepatan dan stabilitas pelatihan.