

BAB II

TINJAUAN PUSATAKA

Dalam Penyusunan Tugas Akhir ini, diperlukan adanya referensi dan tinjauan mengenai penelitian sebelumnya yang memiliki kesamaan topik. Referensi-referensi tersebut nantinya akan digunakan sebagai pembanding sekaligus acuan teori dalam pembuatan tugas akhir dengan topik *deep learning*.

Zhao dkk. Pada tahun 2019 mengusulkan dua arsitektur hibrida berbasis *deep learning* untuk pengenalan emosi suara, yaitu 1D CNN-LSTM dan 2D CNN-LSTM. Model 1D CNN-LSTM digunakan untuk mengekstraksi fitur langsung dari *raw audio*, sedangkan 2D CNN-LSTM memanfaatkan *log-mel spectrogram* sebagai *input*. Kedua arsitektur menggabungkan CNN untuk ekstraksi fitur lokal dan LSTM untuk menangkap pola temporal dalam data suara. Evaluasi dilakukan pada dataset Berlin EmoDB dan IEMOCAP.

Hasil menunjukkan bahwa model 2D CNN-LSTM menghasilkan akurasi terbaik, mencapai 95,89% pada Berlin EmoDB dan 89,16% pada IEMOCAP. Strategi seperti *regularization*, *early stopping*, dan *cross-validation* digunakan untuk mencegah *overfitting*. Penelitian ini membuktikan bahwa kombinasi CNN dan LSTM, khususnya dengan *input log-mel spectrogram*, sangat efektif dalam meningkatkan akurasi pengenalan emosi suara [8].

Li dkk. Pada tahun 2023 mengusulkan pendekatan baru untuk pengenalan emosi suara menggunakan arsitektur 1D Convolutional Neural Network (CNN) tanpa mekanisme *attention*. Pendekatan ini menggabungkan tiga jenis fitur *audio*, yaitu *Melspectrogram*, *Mel Frequency Cepstral Coefficients* (MFCC), dan *Log-Melspectrogram*, untuk menangkap informasi emosional secara lebih menyeluruh. Arsitektur model terdiri dari enam lapis konvolusi 1D yang diikuti oleh lapisan *dense* dan *dropout*. Model ini dievaluasi menggunakan tiga *dataset benchmark*, yaitu Emo-DB, RAVDESS, dan IEMOCAP.

Hasil eksperimen menunjukkan bahwa metode ini mampu mengungguli beberapa model dasar, dengan peningkatan performa khususnya saat *Log-Melspectrogram* digunakan sebagai fitur tambahan. Model mencapai akurasi

sebesar 88,07% pada Emo-DB, 75,25% pada RAVDESS, dan 65,35% pada IEMOCAP, serta mencatat nilai *unweighted average recall* (UAR) yang kompetitif. Studi ini membuktikan bahwa 1D CNN dengan kombinasi fitur pelengkap dapat menjadi solusi yang efisien dan akurat dalam tugas pengenalan emosi suara [9].

Wani dkk. Pada tahun 2020 melakukan penelitian untuk meningkatkan kinerja pengenalan emosi suara (*Speech Emotion Recognition/SER*) dengan membandingkan arsitektur Convolutional Neural Network (CNN) standar dengan Deep Stride Convolutional Neural Network (DSCNN). Studi ini menggunakan input berupa *spectrogram* dari dataset SAVEE dan berfokus pada empat kategori emosi, yaitu marah, senang, netral, dan sedih. Arsitektur DSCNN menggantikan *pooling layer* dengan pendekatan *stride-based* untuk mereduksi dimensi data. Hasilnya menunjukkan bahwa DSCNN mampu mencapai akurasi sebesar 87,8%, lebih tinggi dibandingkan CNN standar yang hanya mencapai 79,4%. Penelitian ini menunjukkan bahwa modifikasi arsitektur CNN dengan strategi *stride* tidak hanya meningkatkan akurasi klasifikasi emosi, tetapi juga memperbaiki efisiensi komputasi, sehingga cocok untuk diterapkan dalam sistem SER yang lebih optimal [10].

Penelitian oleh Yao dkk. pada tahun 2020 mengusulkan kerangka kerja baru untuk pengenalan emosi suara dengan mengintegrasikan tiga jenis model jaringan saraf, yaitu *Deep Neural Network* (DNN) dengan fungsi statistik tingkat tinggi (*High-Level Statistical Functions/HSFs*), *Convolutional Neural Network* (CNN) dengan masukan *mel-spectrogram*, serta *Recurrent Neural Network* (RNN) dengan deskriptor tingkat rendah (*Low-Level Descriptors/LLDs*). Sistem ini dirancang untuk mengenali empat kategori emosi, yakni marah, senang, netral, dan sedih, menggunakan dataset IEMOCAP.

Salah satu inovasi utama dalam studi ini adalah penggunaan mekanisme *attention* berbasis *weighted pooling* pada CNN dan RNN, yang memungkinkan model untuk lebih fokus pada bagian ucapan yang mengandung informasi emosional penting. Selain itu, pendekatan *multi-task learning* digunakan untuk

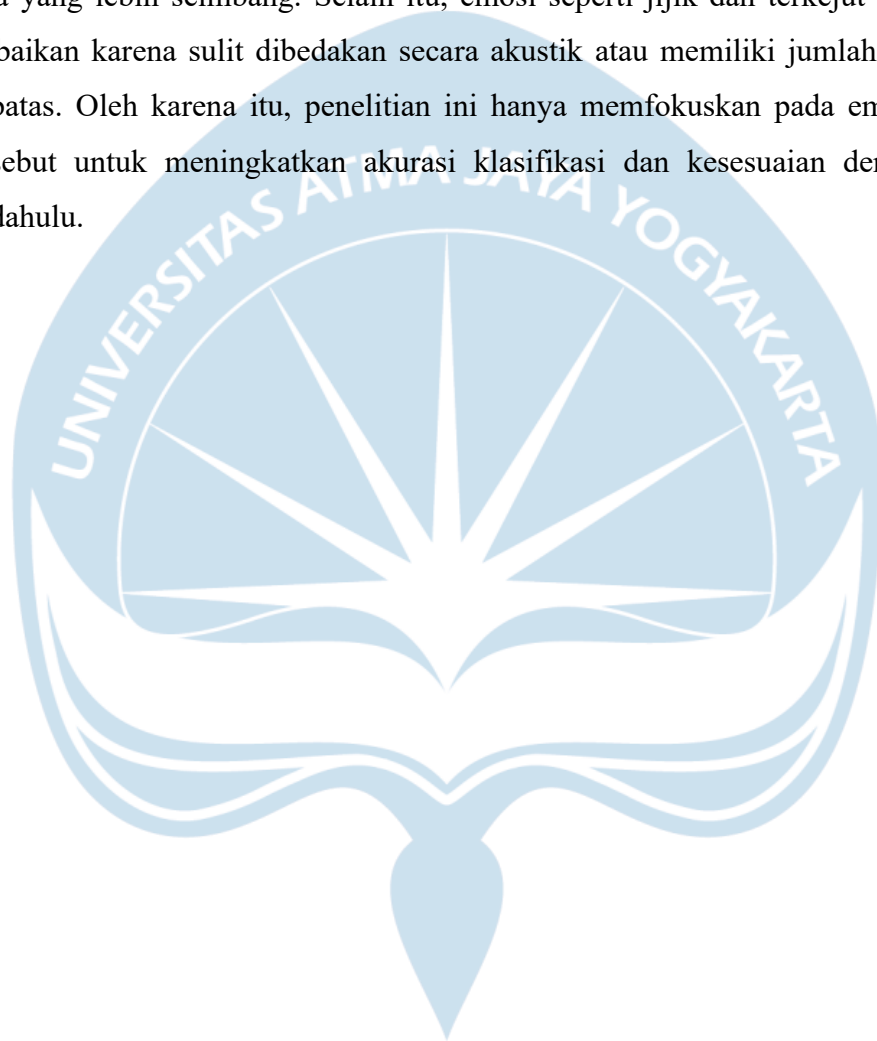
melakukan klasifikasi emosi secara kategorikal sekaligus melakukan regresi terhadap atribut emosi kontinu seperti *valence* dan *activation*, sehingga menghasilkan representasi fitur yang lebih umum dan kuat.

Untuk meningkatkan akurasi, Yao dkk. juga mengembangkan strategi *decision-level fusion* berbasis kepercayaan, yang menggabungkan kekuatan ketiga model secara optimal. Hasil eksperimen menunjukkan bahwa sistem gabungan ini menghasilkan *weighted accuracy* sebesar 57,1% dan akurasi tidak tertimbang sebesar 58,3%, lebih baik dibandingkan masing-masing model secara individual maupun baseline SVM tradisional. Studi ini menunjukkan bahwa kombinasi antara *attention mechanism*, *multi-task learning*, dan *fusion strategy* memberikan kontribusi signifikan dalam meningkatkan performa sistem pengenalan emosi suara [11].

Krishnan dkk. Pada tahun 2021 mengusulkan pendekatan baru dalam pengenalan emosi suara dengan memanfaatkan *Empirical Mode Decomposition* (EMD) dan fitur berbasis entropi non-linear. Dalam studi ini, sinyal suara dari dataset TESS diuraikan menjadi beberapa *Intrinsic Mode Functions* (IMF), yang kemudian dikategorikan menjadi tiga kelompok frekuensi: tinggi, sedang, dan rendah. Dari masing-masing kelompok, diekstraksi lima jenis entropi, yaitu *approximate entropy*, *sample entropy*, *SVD entropy*, *permutation entropy*, dan *spectral entropy*, untuk membentuk vektor fitur komposit.

Vektor fitur ini digunakan untuk melatih berbagai model klasifikasi, seperti LDA, Naïve Bayes, K-NN, SVM, Random Forest, dan Gradient Boosting Machine. Berdasarkan evaluasi menggunakan *tenfold cross-validation*, model LDA menunjukkan performa terbaik dengan akurasi seimbang sebesar 93,3%, skor F1 sebesar 87,9%, dan AUC mencapai 0,995. Hasil ini menunjukkan bahwa kombinasi EMD dan fitur entropi non-linear mampu mengenali tujuh jenis emosi—sedih, marah, jijik, senang, terkejut, menyenangkan, dan netral—dengan tingkat akurasi tinggi, serta melampaui metode konvensional lainnya. Penulis juga mencatat bahwa kualitas data TESS yang tinggi membuat proses praproses tambahan tidak diperlukan [12].

Selain itu beberapa penelitian terdahulu [13], [14], [15], [16] mengenai klasifikasi emosi atau *Speech Emotion Recognition* yang menggunakan deep learning, hanya menggunakan 4 emosi saja yaitu: marah, senang sedih dan netral. karena keempat emosi ini memiliki ciri akustik yang lebih jelas serta representasi data yang lebih seimbang. Selain itu, emosi seperti jijik dan terkejut sering kali diabaikan karena sulit dibedakan secara akustik atau memiliki jumlah data yang terbatas. Oleh karena itu, penelitian ini hanya memfokuskan pada empat emosi tersebut untuk meningkatkan akurasi klasifikasi dan kesesuaian dengan studi terdahulu.



Tabel 2. 1 Tabel Perbandingan Penelitian

Peneliti	Zhao dkk.	Li dkk.	Wani dkk.	Yao dkk.	Krishnan dkk.	William
Metode Preprocessing	<i>Audio Clip Preparation, Log-mel Spectrogram Extraction</i>	<i>Audio Loading and Resampling, FFT Windowing, Melspectrogram, Log-Melspectrogram (LogMels), Mel Frequency Cepstral Coefficients (MFCC) with 40 coefficients returned, Chromagram, Spectral</i>	<i>the short-term Fourier transform (STFT)</i>	<i>Sampling, Framing, Mel-Spectrogram, Low-Level Descriptors (LLD), High-level Statistical Functions (HSF)</i>	<i>approximate entropy, sample entropy, SVD entropy, permutation entropy, dan spectral entropy</i>	<i>Mel Frequency Cepstral Coefficients (MFCC) with 40 coefficients returned, Root Mean Square Energy (RMSE), Zero Crossing Rate (ZCR), Spectral Centroid, Spectral Rolloff, Spectral Bandwidth, Chroma STFT, Data</i>

		<i>contrast</i> <i>Tonal centroid</i> <i>features</i>				<i>Augmentation</i> <i>(noise, pitch shift,</i> <i>stretch, change</i> <i>speed, time shift)</i>
matriks evaluasi	<i>recognition accuracy</i>	<i>Accuracy,</i> <i>precision,</i> <i>Recall, F1-</i> <i>score, Weighted</i> <i>Average Recall</i> <i>(WAR),</i> <i>Unweighted</i> <i>Average Recall</i> <i>(UAR), Macro</i> <i>Average</i> <i>Precision</i> <i>(MAP),</i> <i>Confusion</i> <i>Matrix</i>	<i>prediction</i> <i>accuracy</i>	<i>Weighted</i> <i>Accuracy (WA),</i> <i>Unweighted</i> <i>Accuracy (UA)</i>	<i>Balanced</i> <i>accuracy</i> <i>,F1-score,</i> <i>Recall,</i> <i>Area under the</i> <i>curve (AUC),</i> <i>Specificity,</i> <i>Precision,</i> <i>tenfold cross-</i> <i>validation</i>	<i>Accuracy, Loss,</i> <i>Confusion Matrix,</i> <i>Precision, Recall,</i> <i>F1-Score, Support,</i> <i>Probability Output</i>

keterangan dimensi data	1D & 2D	1D	2D	1D & 2D	1D	1D
Dataset yang digunakan	<i>open-source Berlin Database of Emotional Speech (Emo-DB) dan The Interactive Emotional Dyadic Motion Capture (IEMOCAP)</i>	<i>open-source Berlin Database of Emotional Speech (Emo-DB) , Ravdess (Ryerson Audio-visual database of Emotional Speech and Song) dan The Interactive Emotional Dyadic Motion Capture (IEMOCAP)</i>	<i>Surrey Audio-Visual Expressed Emotion (SAVEE)</i>	<i>The Interactive Emotional Dyadic Motion Capture (IEMOCAP)</i>	<i>Toronto Emotional Speech Set (TESS)</i>	<i>Ryerson Audio-visual database of Emotional Speech and Song (Ravdess), Crema-D, Surrey Audio-Visual Expressed Emotion (Savee), Toronto Emotional Speech Set (Tess), Interactive Emotional Dyadic Motion Capture (IEMOCAP)</i>

Model	<i>convolutional neural network (CNN) dan Long Short-Term Memory (LSTM)</i>	<i>convolutional neural network (CNN)</i>	<i>convolutional neural network (CNN) dan neural network (DSCNN)</i>	HSF-DNN, MS-CNN dan LLD-RNN	<i>Linear Discriminant Analysis (LDA), Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest, dan Gradient Boosting Machine</i>	<i>convolutional neural network (CNN), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), Gated Recurrent Unit (GRU)</i>
--------------	---	---	--	-----------------------------	--	---

Pada tabel 2.1 dapat terlihat jika penelitian ini menggunakan pendekatan yang lebih komprehensif dibanding lima penelitian sebelumnya. Dalam tahap *preprocessing*, penelitian ini menggunakan kombinasi fitur MFCC, RMSE, ZCR, dan fitur spektral, serta menerapkan *data augmentation*, yang tidak umum digunakan dalam penelitian lain. Dari sisi dataset, penelitian ini menggabungkan lima dataset populer (RAVDESS, CREMA-D, SAVEE, TESS, dan IEMOCAP sebagai *data testing*), sedangkan penelitian lain umumnya hanya menggunakan satu atau dua. Untuk model, penelitian ini memakai kombinasi CNN, LSTM, BiLSTM, dan GRU, yang lebih variatif dibandingkan model-model sebelumnya. Evaluasi juga dilakukan lebih menyeluruh dengan metrik tambahan seperti *loss*, *probabilitas output*, dan *support*. Hal ini menunjukkan bahwa pendekatan pada penelitian ini lebih kuat dalam hal generalisasi dan ketepatan klasifikasi emosi dari suara.