KOMPARASI CNN-BILSTM, CRNN-BIGRU-ATTENTION, DAN EMOTION2VEC+ UNTUK DETEKSI EMOSI SUARA PELANGGAN

Muhammad Revin Putra Fahlefi[1], Bayu Sulistiyanto Ipung Sutejo[2], Riya Widayanti[3], Adi Widiantono[4]
Program Studi Teknik Informatika, Fakultas Ilmu Komputer
Universitas Esa Unggul[1,2,3,4]
Jalan Arjuna Utara No 9, Kebon Jeruk, Jakarta Barat
revinfahlefi@student.esaunggul.ac.id[1], bayu.sulistiyanto@esaunggul.ac.id[2],
riya.widayanti@esaunggul.ac.id[3], adi.widiantono@esaunggul.ac.id[4]

Abstrak

Pengenalan emosi suara dapat mendukung layanan pelanggan dalam memahami kondisi emosional pengguna secara lebih cepat dan konsisten. Penelitian ini membandingkan CNN-BiLSTM, CRNN-BiGRU-Attention, dan emotion2vec+ untuk mengklasifikasikan anger, sadness, happy, dan neutral. Data terdiri atas 5.764 rekaman terpilih dari RAVDESS dan CREMA-D yang diproses secara terpisah menggunakan speaker-independent split. CNN-BiLSTM dan CRNN-BiGRU-Attention dilatih dari awal menggunakan log-mel spectrogram, sedangkan emotion2vec+ melakukan inferensi langsung pada raw waveform tanpa fine-tuning. Evaluasi mencakup data bersih, *babble noise* dan *white noise* pada SNR 15, 10, dan 5 dB, pengujian *cross-corpus*, *false negative rate* (FNR) pada anger dan sadness, serta efisiensi komputasi. emotion2vec+ memperoleh Unweighted Average Recall (UAR) data bersih tertinggi sebesar 93,49% pada RAVDESS dan 91,52% pada CREMA-D, rata-rata UAR kondisi *noise* sebesar 82,68% dan 71,67%, serta maksimum FNR emosi kritis sebesar 9,38% dan 14,29%. UAR *cross-corpus* tertinggi hanya mencapai 33,69%, yang menunjukkan keterbatasan generalisasi antarkorpus. CRNN-BiGRU-Attention menjadi model paling efisien dengan 1.692.997 parameter, ukuran bobot 6,47 MB, dan waktu inferensi 1,3805-1,6569 milidetik per sampel. Dengan demikian, emotion2vec+ unggul dalam kinerja klasifikasi dan ketahanan terhadap gangguan, sedangkan CRNN-BiGRU-Attention lebih sesuai ketika efisiensi komputasi diprioritaskan.

Kata Kunci: *Speech Emotion Recognition*, CNN-BiLSTM, CRNN-BiGRU-Attention, emotion2vec+, layanan pelanggan

Abstract

Speech emotion recognition can support customer service by identifying users' emotional states more quickly and consistently. This study compares CNN-BiLSTM, CRNN-BiGRU-Attention, and emotion2vec+ in classifying anger, sadness, happy, and neutral. The data comprise 5,764 selected recordings from RAVDESS and CREMA-D, processed separately using a speaker-independent split. CNN-BiLSTM and CRNN-BiGRU-Attention were trained from scratch using log-mel spectrograms, whereas emotion2vec+ performed direct inference on raw waveforms without fine-tuning. The evaluation covered clean audio, babble noise and white noise at SNR levels of 15, 10, and 5 dB, cross-corpus testing, false negative rate (FNR) for anger and sadness, and computational efficiency. emotion2vec+ achieved the highest clean-data Unweighted Average Recall (UAR), reaching 93.49% on RAVDESS and 91.52% on CREMA-D, average noisy-condition UARs of 82.68% and 71.67%, and maximum critical-emotion FNRs of 9.38% and 14.29%, respectively. The highest cross-corpus UAR was only 33.69%, indicating limited cross-corpus generalization. CRNN-BiGRU-Attention was the most efficient model, with 1,692,997 parameters, a 6.47 MB weight file, and an inference time of 1.3805-1.6569 milliseconds per sample. Thus, emotion2vec+ offered superior classification performance and noise robustness, whereas CRNN-BiGRU-Attention was preferable when computational efficiency was prioritized.

***Keywords**: Speech Emotion Recognition, CNN-BiLSTM, CRNN-BiGRU-Attention, emotion2vec+, customer service*

PENDAHULUAN

Layanan pelanggan berbasis suara memungkinkan pelanggan menyampaikan pertanyaan, kebutuhan, dan keluhan secara langsung. Informasi dalam percakapan tidak hanya berasal dari kata-kata, tetapi juga dari intonasi, energi, tempo, dan pola pengucapan yang dapat mencerminkan kondisi emosional pembicara. Kemampuan mengenali emosi secara otomatis dapat membantu pemantauan interaksi, terutama ketika pelanggan menunjukkan kemarahan atau kesedihan yang memerlukan perhatian lebih lanjut. Namun, penerapan pada lingkungan nyata menghadapi variasi karakteristik pembicara,

perangkat perekaman, dan gangguan suara. Parra-Gallego dan Orozco-Arroyave (2022) menunjukkan bahwa kondisi akustik yang tidak terkontrol dapat memengaruhi pengenalan emosi dan evaluasi kepuasan pelanggan melalui suara.

Pemrosesan emosi dari sinyal suara dikenal sebagai *Speech Emotion Recognition* (SER). Perkembangan *deep learning* menghasilkan berbagai arsitektur yang mampu mempelajari pola waktu-frekuensi dan hubungan temporal pada audio. Zhao et al. (2019) menunjukkan bahwa kombinasi CNN dan LSTM dapat memanfaatkan CNN untuk mengekstraksi pola lokal, sedangkan LSTM digunakan untuk mempelajari hubungan temporal. Penggunaan jaringan dua arah dan mekanisme *attention* dapat memperluas konteks yang dipelajari serta memberikan bobot lebih besar pada bagian audio yang relevan terhadap emosi.

Pendekatan lain berkembang melalui model *pretrained* berbasis *self-supervised learning*. Ma et al. (2024) memperkenalkan emotion2vec sebagai representasi emosi suara yang dipelajari dari data berskala besar. Model *pretrained* berpotensi memberikan kinerja yang lebih baik dibandingkan model yang dilatih dari awal pada data terbatas, tetapi umumnya memiliki ukuran dan kebutuhan komputasi yang lebih besar. Oleh karena itu, pemilihan model tidak cukup hanya didasarkan pada satu nilai klasifikasi. Sutejo et al. (2025) juga menunjukkan bahwa penilaian sistem digital dapat dilakukan secara lebih terarah melalui penggunaan beberapa indikator terukur. Prinsip tersebut diterapkan dalam penelitian ini dengan mempertimbangkan kinerja klasifikasi, ketahanan terhadap gangguan, deteksi emosi kritis, dan efisiensi komputasi.

Penelitian terdahulu masih banyak mengevaluasi model pada data bersih atau satu korpus tertentu. Hasil tersebut belum sepenuhnya menggambarkan kemampuan model ketika menghadapi gangguan suara maupun sumber data yang berbeda. Selain itu, nilai *accuracy* secara umum belum menunjukkan seberapa sering model gagal mengenali emosi negatif. Dalam konteks layanan pelanggan, kesalahan pada kelas *anger* dan *sadness* perlu diperhatikan karena dapat menyebabkan kondisi pelanggan yang membutuhkan tindak lanjut tidak teridentifikasi.

Berdasarkan permasalahan tersebut, penelitian ini membandingkan CNN-BiLSTM, CRNN-BiGRU-Attention, dan emotion2vec+ menggunakan RAVDESS dan CREMA-D. Evaluasi dilakukan pada data bersih, kondisi *babble noise* dan *white noise*, pengujian *cross-corpus*, *false negative rate* pada kelas *anger* dan *sadness*, serta efisiensi komputasi. Penelitian ini bertujuan mengetahui keunggulan dan keterbatasan setiap model serta menentukan model utama berdasarkan kinerja klasifikasi, ketahanan terhadap gangguan suara, kemampuan mendeteksi emosi kritis, dan efisiensi komputasi.

PENELITIAN RELEVAN

Zhao et al. (2019) mengembangkan model CNN-LSTM untuk mengenali emosi dari sinyal suara dan *log-mel spectrogram*. CNN digunakan untuk mengekstraksi pola lokal, sedangkan LSTM mempelajari hubungan temporal pada fitur yang dihasilkan. Penelitian tersebut menunjukkan bahwa penggabungan CNN dan jaringan rekuren mampu memberikan hasil yang lebih baik dibandingkan penggunaan CNN tanpa pemodelan temporal. Namun, arsitektur yang digunakan belum memanfaatkan pemrosesan dua arah dan mekanisme *attention*.

Pengembangan berikutnya dilakukan oleh Liu et al. (2023) melalui kombinasi CNN, *bidirectional LSTM*, *attention*, dan *multi-task learning*. Penggunaan jaringan dua arah memungkinkan model mempelajari konteks sebelum dan sesudah suatu bagian suara, sedangkan *attention* memberikan bobot lebih besar pada bagian yang dianggap penting. Hasil penelitian tersebut memperlihatkan bahwa pemodelan temporal dan pemberian bobot perhatian dapat meningkatkan kemampuan generalisasi SER, tetapi membutuhkan representasi masukan dan prosedur pelatihan yang lebih kompleks.

Pendekatan berbasis model pra-latih dikembangkan oleh Ma et al. (2024) melalui emotion2vec. Model ini mempelajari representasi emosi menggunakan *self-supervised online distillation* dengan menggabungkan pembelajaran tingkat bingkai dan tingkat ujaran. emotion2vec+ menunjukkan hasil yang konsisten pada berbagai dataset dan bahasa, sehingga dapat digunakan sebagai representasi umum untuk tugas pengenalan emosi suara. Berbeda dari penelitian tersebut yang mengevaluasi representasi melalui pelatihan lapisan klasifikasi, penelitian ini menggunakan varian emotion2vec+ untuk inferensi langsung tanpa *fine-tuning* pada RAVDESS dan CREMA-D.

Ketahanan terhadap kondisi nyata juga menjadi perhatian dalam penelitian SER. Leem et al. (2021) menunjukkan bahwa gangguan nonstasioner dan perubahan SNR dapat menurunkan kinerja

pengenalan emosi, sehingga evaluasi pada audio bersih belum cukup untuk menggambarkan ketahanan model. Sementara itu, Parra-Gallego dan Orozco-Arroyave (2022) menemukan bahwa fitur yang efektif pada dataset emosi terkontrol belum tentu memberikan hasil terbaik pada rekaman layanan pelanggan dengan kondisi akustik nyata.

Aftab et al. (2022) mengembangkan LIGHT-SERNET sebagai model konvolusional ringan untuk pengenalan emosi suara, tetapi belum membandingkannya dengan model pra-latih dalam berbagai kondisi pengujian.

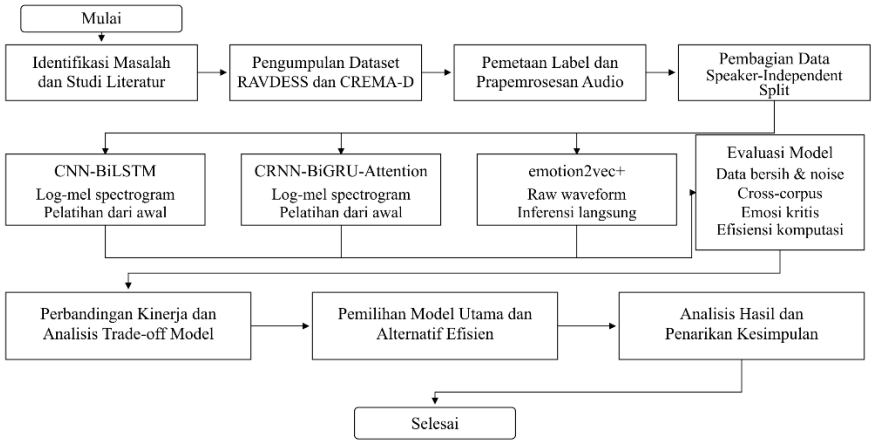
Pepino et al. (2021) menunjukkan bahwa representasi wav2vec 2.0 dapat dimanfaatkan untuk pengenalan emosi suara, tetapi belum mengevaluasi ketahanan, generalisasi lintas korpus, dan efisiensi komputasi secara bersamaan.

Penelitian terdahulu telah membahas arsitektur CNN-rekuren, mekanisme attention, model pra-latih, dan ketahanan terhadap gangguan suara melalui fokus evaluasi yang berbeda. Namun, pada penelitian yang ditinjau, aspek kinerja data bersih, ketahanan terhadap *noise*, generalisasi *cross-corpus*, kegagalan mendeteksi emosi kritis, dan efisiensi komputasi belum dibandingkan secara bersamaan dalam satu kerangka evaluasi. Penelitian ini mengisi kesenjangan tersebut dengan membandingkan CNN-BiLSTM, CRNN-BiGRU-Attention, dan emotion2vec+ pada RAVDESS dan CREMA-D menggunakan skenario serta metrik evaluasi yang telah ditetapkan.

METODE PENELITIAN

Desain Penelitian

Penelitian ini menggunakan metode eksperimen komparatif untuk membandingkan kinerja CNN-BiLSTM, CRNN-BiGRU-Attention, dan emotion2vec+ dalam mengklasifikasikan empat kelas emosi, yaitu anger, sadness, happy, dan neutral. Tahapan penelitian meliputi pengumpulan dataset, pemetaan label, prapemrosesan audio, pembagian data berdasarkan pembicara, pelatihan atau inferensi model, pengujian pada beberapa skenario, serta analisis trade-off antara kinerja dan efisiensi komputasi.



Gambar 1. Alur Penelitian

Data dan Pemrosesan

Data penelitian dikumpulkan melalui studi dokumentasi dengan menggunakan rekaman audio dari dataset RAVDESS (Livingstone & Russo, 2018) dan CREMA-D (Cao et al., 2014). Dari kedua dataset tersebut dipilih empat kelas emosi yang tersedia secara bersama, yaitu anger, sadness, happy, dan neutral. Setelah proses seleksi dan pemetaan label, digunakan 864 rekaman RAVDESS dan 4.900 rekaman CREMA-D. Kedua dataset diproses dan digunakan secara terpisah serta tidak digabungkan menjadi satu kelompok data pelatihan. Pembagian data dilakukan menggunakan speaker-independent split agar rekaman dari pembicara yang sama tidak muncul pada lebih dari satu kelompok data. Rincian pembagian data ditampilkan pada Tabel 1.

Tabel 1. Pembagian Dataset Penelitian

No	Dataset	Pelatihan	Validasi	Pengujian	Total
1	RAVDESS	576	144	144	864
2	CREMA-D	3.392	752	756	4.900
	Total	3.968	896	900	5.764

Seluruh audio dikonversi menjadi satu kanal (mono), diseragamkan pada sample rate 16.000 Hz, dinormalisasi amplitudonya, dan disesuaikan menjadi durasi tiga detik. Audio yang berdurasi lebih dari tiga detik dipotong, sedangkan audio yang lebih pendek dilengkapi menggunakan zero-padding. Proses tersebut menghasilkan raw waveform sepanjang 48.000 sampel yang digunakan sebagai masukan emotion2vec+.

Untuk CNN-BiLSTM dan CRNN-BiGRU-Attention, raw waveform diubah menjadi log-mel spectrogram menggunakan 128 mel bin, ukuran FFT 400, dan hop length 160. Proses ekstraksi tersebut menghasilkan fitur dengan bentuk [1, 128, 301]. Seluruh 5.764 rekaman audio diproses menggunakan konfigurasi prapemrosesan yang sama.

Model dan Prosedur Pelatihan

CNN-BiLSTM dan CRNN-BiGRU-Attention menggunakan log-mel spectrogram berbentuk [1, 128, 301] sebagai masukan. Kedua model menggunakan dua blok konvolusi yang terdiri atas Conv2D, BatchNorm2D, ReLU, MaxPool2D, dan dropout. Jumlah kanal keluaran pada kedua blok konvolusi berturut-turut adalah 32 dan 64 dengan ukuran kernel 3 × 3. Hasil ekstraksi CNN kemudian diubah menjadi urutan temporal sebelum diteruskan ke lapisan rekuren.

Pada CNN-BiLSTM, urutan temporal diproses menggunakan BiLSTM dengan hidden size 128 dan dua arah pemrosesan. Representasi dari kedua arah menghasilkan vektor berukuran 256 yang diteruskan ke dropout dan lapisan linear untuk menghasilkan empat kelas emosi.

Pada CRNN-BiGRU-Attention, urutan temporal diproses menggunakan BiGRU dengan hidden size 128 dan dua arah pemrosesan. Seluruh keluaran temporal kemudian diberi bobot menggunakan mekanisme attention untuk menghasilkan vektor konteks berukuran 256. Vektor tersebut diteruskan ke dropout dan lapisan linear untuk menghasilkan empat kelas emosi. CNN-BiLSTM memiliki 2.250.308 parameter, sedangkan CRNN-BiGRU-Attention memiliki 1.692.997 parameter.

Kedua model dilatih dari awal secara terpisah pada RAVDESS dan CREMA-D. Pelatihan menggunakan CrossEntropyLoss, optimizer AdamW, learning rate awal 0,001, weight decay 0,0001, batch size 32, dan maksimum 50 epoch. ReduceLROnPlateau digunakan untuk menurunkan learning rate dengan faktor 0,5 apabila UAR validasi tidak meningkat selama tiga epoch. Gradient clipping diterapkan dengan norma maksimum 5,0. Checkpoint terbaik dipilih berdasarkan UAR validasi tertinggi, sedangkan early stopping menghentikan pelatihan apabila tidak terjadi peningkatan selama sepuluh epoch.

emotion2vec+ menggunakan model iic/emotion2vec_plus_base yang dimuat melalui FunASR dan ModelScope. Model menerima raw waveform sepanjang 48.000 sampel dan menjalankan inferensi menggunakan fungsi generate() pada tingkat utterance. Label keluaran model dibersihkan dan dipetakan ke kelas anger, sadness, happy, neutral, atau other. Label other tidak dipaksakan menjadi salah satu dari empat kelas target penelitian.

Skenario Evaluasi

Pengujian pada data bersih dilakukan menggunakan kelompok data pengujian dari masing-masing dataset tanpa penambahan gangguan suara. Kinerja ketiga model diukur menggunakan accuracy, Unweighted Average Recall (UAR), Macro-F1, dan confusion matrix.

Ketahanan model diuji dengan menambahkan *babble noise* dan *white noise* pada data pengujian. Masing-masing jenis gangguan diberikan pada tingkat SNR 15 dB, 10 dB, dan 5 dB. Dengan demikian, setiap kombinasi model dan dataset diuji pada enam kondisi gangguan. Data yang telah diberi gangguan hanya digunakan untuk pengujian dan tidak digunakan selama proses pelatihan.

Pengujian *cross-corpus* dilakukan pada CNN-BiLSTM dan CRNN-BiGRU-Attention dalam dua arah, yaitu model yang dilatih pada RAVDESS diuji menggunakan CREMA-D dan model yang dilatih pada CREMA-D diuji menggunakan RAVDESS. emotion2vec+ tidak disertakan dalam pengujian *cross-corpus* karena model tersebut tidak dilatih secara khusus menggunakan salah satu korpus penelitian.

Kemampuan model dalam mendeteksi emosi kritis dianalisis menggunakan *false negative rate* (FNR) pada kelas anger dan sadness. FNR dihitung dengan membagi jumlah data suatu kelas yang salah diprediksi dengan seluruh data aktual pada kelas tersebut. Nilai FNR yang lebih rendah menunjukkan bahwa model lebih jarang melewatkan kelas emosi yang dievaluasi. Perbandingan antar-model diringkas menggunakan nilai FNR maksimum antara kelas anger dan sadness pada setiap kombinasi model dan dataset.

Efisiensi komputasi dievaluasi berdasarkan jumlah parameter, ukuran berkas model, puncak penggunaan memori GPU, waktu inferensi per sampel, dan *throughput*.

Pengukuran CNN-BiLSTM dan CRNN-BiGRU-Attention dilakukan menggunakan fitur log-mel spectrogram yang telah tersedia dan hanya mencakup proses forward model. Pengukuran emotion2vec+ dilakukan melalui fungsi inferensi bawaan menggunakan berkas WAV. Oleh karena itu, hasil efisiensi digunakan untuk menggambarkan kebutuhan komputasi pada jalur inferensi masing-masing model dan tidak diposisikan sebagai perbandingan end-to-end yang sepenuhnya identik.

HASIL DAN PEMBAHASAN

Kinerja pada Data Bersih

Pengujian data bersih digunakan untuk mengetahui kemampuan dasar setiap model sebelum diberi gangguan suara. Hasil pengujian ditampilkan pada Tabel 2.

Tabel 2. Kinerja Model pada Data Bersih

No	Model	Dataset	*Accuracy*	UAR	Macro-F1
1	CNN-BiLSTM	RAVDESS	59,03	57,55	55,90
2	CNN-BiLSTM	CREMA-D	62,04	61,78	62,01
3	CRNN-BiGRU-Attention	RAVDESS	68,06	68,49	68,33
4	CRNN-BiGRU-Attention	CREMA-D	70,37	70,32	70,39
5	emotion2vec+	RAVDESS	93,75	93,49	94,16
6	emotion2vec+	CREMA-D	91,27	91,52	91,30

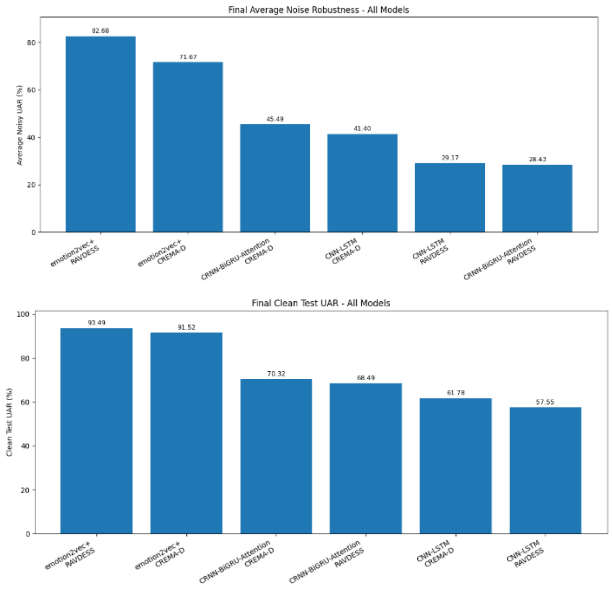
emotion2vec+ menghasilkan kinerja tertinggi pada seluruh metrik dan kedua dataset. Pada RAVDESS, emotion2vec+ memperoleh UAR sebesar 93,49%, lebih tinggi 25,00 poin persentase daripada CRNN-BiGRU-Attention dan 35,94 poin persentase daripada CNN-BiLSTM. Pada CREMA-D, emotion2vec+ memperoleh UAR sebesar 91,52%, dengan selisih masing-masing 21,20 dan 29,74 poin persentase dibandingkan CRNN-BiGRU-Attention dan CNN-BiLSTM.

Hasil tersebut menunjukkan bahwa representasi emosi pada emotion2vec+ dapat digunakan secara efektif untuk mengklasifikasikan empat kelas emosi pada kedua dataset penelitian. Meskipun digunakan melalui inferensi langsung, model tersebut tetap menghasilkan kinerja lebih tinggi daripada kedua model yang dilatih dari awal.

Di antara model yang dilatih dari awal, CRNN-BiGRU-Attention menghasilkan kinerja lebih tinggi daripada CNN-BiLSTM pada kedua dataset. BiGRU mempelajari hubungan temporal dari dua arah, sedangkan mekanisme attention memberikan bobot yang berbeda pada setiap bagian urutan sehingga model dapat memusatkan representasi pada informasi temporal yang lebih relevan. Temuan ini sejalan dengan Liu et al. (2023), yang menunjukkan bahwa pemrosesan dua arah dan mekanisme perhatian dapat mendukung kinerja pengenalan emosi suara.

Ketahanan terhadap Gangguan Suara

Ketahanan model dievaluasi menggunakan *babble noise* dan *white noise* pada tingkat SNR 15 dB, 10 dB, dan 5 dB. Setiap kombinasi model dan dataset diuji pada enam kondisi gangguan. Rata-rata UAR dihitung dari UAR pada seluruh kombinasi jenis gangguan dan tingkat SNR tersebut. Perbandingan antara UAR data bersih dan rata-rata UAR pada kondisi gangguan ditampilkan pada Gambar 2.



Gambar 2. Perbandingan UAR Data Bersih dan Rata-Rata UAR pada Kondisi *Noise*

emotion2vec+ mempertahankan rata-rata UAR tertinggi pada kedua dataset, yaitu 82,68% pada RAVDESS dan 71,67% pada CREMA-D. CRNN-BiGRU-Attention menghasilkan rata-rata UAR sebesar 28,43% pada RAVDESS dan 45,49% pada CREMA-D, sedangkan CNN-BiLSTM menghasilkan rata-rata UAR sebesar 29,17% dan 41,40%.

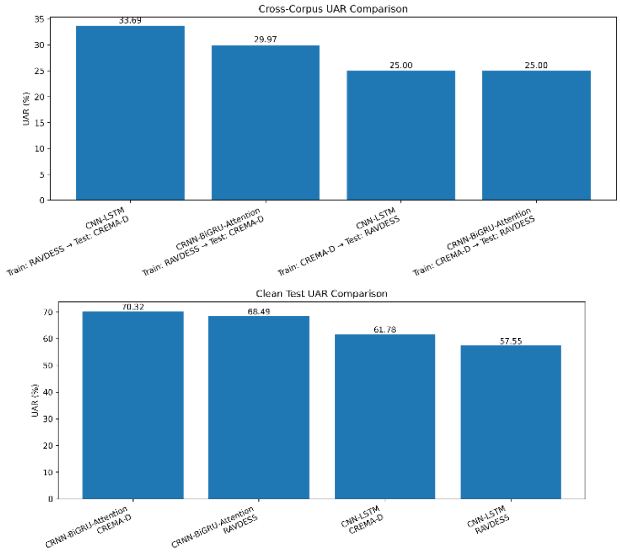
Pada RAVDESS, CNN-BiLSTM memperoleh rata-rata UAR sedikit lebih tinggi daripada CRNN-BiGRU-Attention dengan selisih 0,74 poin persentase. Sebaliknya, pada CREMA-D, CRNN-BiGRU-Attention lebih tinggi 4,09 poin persentase daripada CNN-BiLSTM. Dengan demikian, tidak terdapat satu model yang dilatih dari awal yang selalu lebih baik pada kedua dataset dalam kondisi gangguan suara.

White noise secara umum menghasilkan UAR yang lebih rendah daripada *babble noise*. Pada kondisi *white noise* 5 dB, emotion2vec+ menghasilkan UAR sebesar 54,17% pada RAVDESS dan 43,35% pada CREMA-D. Pada beberapa kondisi gangguan yang lebih berat, CNN-BiLSTM dan CRNN-BiGRU-Attention menghasilkan UAR mendekati 25%, yaitu tingkat peluang acak pada klasifikasi seimbang dengan empat kelas.

Penurunan kinerja pada tingkat SNR yang lebih rendah menunjukkan bahwa peningkatan kekuatan gangguan mengurangi kemampuan model dalam mempertahankan informasi akustik yang berkaitan dengan emosi. Hasil ini mendukung temuan Leem et al. (2021) bahwa evaluasi pada audio bersih belum cukup untuk menggambarkan ketahanan model pada kondisi bising.

Generalisasi Lintas Korpus

Pengujian *cross-corpus* dilakukan untuk mengetahui kemampuan model dalam menghadapi data yang berasal dari korpus berbeda dari korpus pelatihannya. Model yang telah dilatih pada korpus sumber langsung diuji menggunakan data pengujian dari korpus target tanpa pelatihan ulang. Perbandingan UAR pengujian dalam korpus dan *cross-corpus* ditampilkan pada Gambar 3.



Gambar 3. Perbandingan UAR Dalam Korpus dan *Cross-Corpus*

CNN-BiLSTM yang dilatih menggunakan RAVDESS dan diuji pada CREMA-D memperoleh UAR *cross-corpus* tertinggi sebesar 33,69%. Pada arah pengujian yang sama, CRNN-BiGRU-Attention memperoleh UAR sebesar 29,97%. Ketika model dilatih menggunakan CREMA-D dan diuji pada RAVDESS, CNN-BiLSTM dan CRNN-BiGRU-Attention sama-sama menghasilkan UAR sebesar 25,00%.

Seluruh skenario *cross-corpus* menghasilkan UAR yang lebih rendah daripada pengujian dalam korpus. Penurunan terkecil terjadi pada CNN-BiLSTM yang dilatih pada RAVDESS dan diuji pada CREMA-D. UAR model tersebut menurun dari 57,55% pada pengujian dalam korpus menjadi 33,69% pada pengujian *cross-corpus*, atau turun sebesar 23,86 poin persentase.

Penurunan terbesar terjadi pada CRNN-BiGRU-Attention yang dilatih pada CREMA-D dan diuji pada RAVDESS. UAR model tersebut menurun dari 70,32% menjadi 25,00%, atau turun sebesar 45,32 poin persentase. Nilai UAR 25,00% juga menunjukkan bahwa kinerja model telah mencapai tingkat peluang dasar untuk klasifikasi empat kelas.

Hasil tersebut menunjukkan bahwa CNN-BiLSTM dan CRNN-BiGRU-Attention belum mampu mempertahankan kinerjanya ketika korpus pengujian berbeda dari korpus pelatihan. Perbedaan karakteristik pembicara, gaya ekspresi emosi, distribusi kelas, serta kondisi perekaman dapat menyebabkan pergeseran distribusi data yang belum dapat ditangani dengan baik oleh kedua model.

Deteksi Emosi Kritis

Kemampuan model dalam mendeteksi anger dan sadness dievaluasi menggunakan *false negative rate* (FNR). Perbandingan antar-model diringkas menggunakan nilai FNR tertinggi antara kedua kelas pada setiap kombinasi model dan dataset. Nilai yang lebih rendah menunjukkan bahwa model lebih jarang melewatkan emosi kritis. Hasil evaluasi ditampilkan pada Tabel 3.

Tabel 3. Maksimum FNR pada Emosi Kritis

No	Model	RAVDESS (%)	CREMA-D (%)
1	CNN-BiLSTM	75,00	31,63
2	CRNN-BiGRU-Attention	43,75	29,59
3	emotion2vec+	9,38	14,29

emotion2vec+ menghasilkan maksimum FNR terendah pada kedua dataset. Pada RAVDESS, FNR kelas anger sebesar 3,12% dan sadness sebesar 9,38%. Pada CREMA-D, FNR kelas anger sebesar 8,16% dan sadness sebesar 14,29%. Dengan demikian, nilai maksimum FNR emotion2vec+ adalah 9,38% pada RAVDESS dan 14,29% pada CREMA-D.

Di antara model yang dilatih dari awal, CRNN-BiGRU-Attention menghasilkan maksimum FNR yang lebih rendah daripada CNN-BiLSTM pada kedua dataset. Kelas sadness menjadi kelas dengan

FNR tertinggi pada seluruh kombinasi model dan dataset. Berdasarkan confusion matrix, kesalahan pada kelas sadness umumnya mengarah ke neutral. Dalam konteks layanan pelanggan, kesalahan tersebut dapat menyebabkan kondisi pelanggan yang menunjukkan kesedihan tidak teridentifikasi sebagai interaksi yang memerlukan perhatian.

Efisiensi Komputasi

Efisiensi komputasi dibandingkan berdasarkan jumlah parameter, ukuran berkas model, puncak penggunaan memori GPU, waktu inferensi, dan *throughput*. Hasil pengujian pada RAVDESS dan CREMA-D diringkas dalam bentuk rentang nilai pada Tabel 4.

Tabel 4. Ringkasan Efisiensi Komputasi

No	Model	Parameter	Ukuran model	Puncak memori GPU	Waktu Inferensi Terendah	*Throughput* (sampel/detik)
1	CNN-BiLSTM	2,250,308	8,59 MB	60,07 MB	2,0234-2,0426 ms	489,56-494,23
2	CRNN-BiGRU-Attention	1,692,997	6,47 MB	55,75 MB	1,3805-1,6569 ms	603,52-724,40
3	emotion2vec+	93,797,909	1,066,44 MB	443,29 MB	45,9434-48,7768 ms	20,50-21,77

CRNN-BiGRU-Attention menjadi model paling efisien berdasarkan seluruh indikator yang diukur. Model tersebut memiliki jumlah parameter dan ukuran berkas model paling kecil, menggunakan puncak memori GPU sebesar 55,75 MB, serta menghasilkan waktu inferensi antara 1,3805 dan 1,6569 milidetik per sampel. *Throughput* tertingginya mencapai 724,40 sampel per detik.

emotion2vec+ membutuhkan sumber daya komputasi paling besar dengan 93.797.909 parameter, ukuran berkas model sebesar 1.066,44 MB, dan puncak penggunaan memori GPU sebesar 443,29 MB. Waktu inferensinya berada pada rentang 45,9434-48,7768 milidetik per sampel dengan *throughput* antara 20,50 dan 21,77 sampel per detik.

Perbandingan waktu inferensi perlu dibaca dengan memperhatikan cakupan pengukuran. CNN-BiLSTM dan CRNN-BiGRU-Attention diukur melalui forward pass menggunakan log-mel spectrogram yang telah tersedia, sedangkan emotion2vec+ diukur melalui fungsi generate() menggunakan berkas WAV. Oleh karena itu, hasil tersebut menggambarkan efisiensi jalur inferensi masing-masing model dan bukan perbandingan end-to-end yang sepenuhnya identik.

Pemilihan Model

Pemilihan model dilakukan dengan mempertimbangkan kinerja pada data bersih, ketahanan terhadap gangguan suara, kemampuan mendeteksi emosi kritis, dan efisiensi komputasi. Penilaian dilakukan melalui analisis trade-off antarkriteria dan tidak menggunakan skor gabungan berbobot.

Pendekatan berbasis beberapa indikator juga digunakan oleh Sutejo et al. (2025) untuk menentukan prioritas risiko dalam transformasi digital kesehatan. Meskipun berada pada domain yang berbeda, prinsip serupa diterapkan dalam penelitian ini untuk menentukan model utama berdasarkan beberapa aspek evaluasi.

emotion2vec+ menghasilkan UAR data bersih dan rata-rata UAR kondisi gangguan tertinggi serta maksimum FNR terendah pada kedua dataset. Berdasarkan hasil tersebut, emotion2vec+ dipilih sebagai model utama ketika kinerja klasifikasi, ketahanan terhadap gangguan, dan sensitivitas terhadap emosi kritis menjadi prioritas.

CRNN-BiGRU-Attention menghasilkan kinerja terbaik di antara model yang dilatih dari awal serta memiliki jumlah parameter, ukuran berkas model, penggunaan memori GPU, dan waktu inferensi paling rendah. Model tersebut ditetapkan sebagai alternatif ketika efisiensi komputasi menjadi pertimbangan utama.

SIMPULAN

emotion2vec+ menghasilkan kinerja tertinggi pada pengujian data bersih dengan UAR sebesar 93,49% pada RAVDESS dan 91,52% pada CREMA-D. Model tersebut juga mempertahankan rata-rata UAR tertinggi pada kondisi gangguan suara, yaitu 82,68% pada RAVDESS dan 71,67% pada CREMA-D. Dalam mendeteksi emosi kritis, emotion2vec+ menghasilkan maksimum FNR terendah sebesar 9,38% pada RAVDESS dan 14,29% pada CREMA-D. Hasil tersebut menunjukkan bahwa

emotion2vec+ menjadi model utama ketika kinerja klasifikasi, ketahanan terhadap gangguan suara, dan kemampuan mendeteksi emosi kritis menjadi prioritas.

Di antara model yang dilatih dari awal, CRNN-BiGRU-Attention menghasilkan kinerja klasifikasi terbaik dan kebutuhan komputasi paling rendah. Model ini memiliki 1.692.997 parameter, ukuran berkas model sebesar 6,47 MB, puncak penggunaan memori GPU sebesar 55,75 MB, serta waktu inferensi antara 1,3805 dan 1,6569 milidetik per sampel. Oleh karena itu, CRNN-BiGRU-Attention menjadi alternatif ketika efisiensi komputasi lebih diprioritaskan.

Pengujian *cross-corpus* menghasilkan UAR sebesar 25,00-33,69%, lebih rendah daripada pengujian dalam korpus. Hasil tersebut menunjukkan bahwa kemampuan generalisasi CNN-BiLSTM dan CRNN-BiGRU-Attention terhadap korpus yang berbeda masih terbatas.

Penelitian selanjutnya disarankan menggunakan percakapan layanan pelanggan spontan berbahasa Indonesia, memperluas variasi kelas emosi, pembicara, korpus, kondisi perekaman, serta jenis dan tingkat gangguan suara. Penelitian berikutnya juga dapat mengkaji augmentasi pelatihan, adaptasi domain, normalisasi antarkorpus, atau fine-tuning model pra-latih. Pengukuran efisiensi selanjutnya perlu dilakukan menggunakan prosedur end-to-end yang seragam.

DAFTAR PUSTAKA

Aftab, A., Morsali, A., Ghaemmaghami, S., & Champagne, B. (2022). LIGHT-SERNET: A lightweight fully convolutional neural network for speech emotion recognition. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6912–6916. https://doi.org/10.1109/ICASSP43922.2022.9746679

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, *5*(4), 377–390. https://doi.org/10.1109/TAFFC.2014.2336244

Leem, S.-G., Fulford, D., Onnela, J.-P., Gard, D., & Busso, C. (2021). Separation of emotional and reconstruction embeddings on ladder network to improve speech emotion recognition robustness in noisy conditions. *Interspeech 2021*, 2871–2875. https://doi.org/10.21437/Interspeech.2021-1438

Liu, Z.-T., Han, M.-T., Wu, B.-H., & Rehman, A. (2023). Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning. *Applied Acoustics*, *202*, 109178. https://doi.org/10.1016/j.apacoust.2022.109178

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, *13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., & Chen, X. (2024). emotion2vec: Self-supervised pre-training for speech emotion representation. *Findings of the Association for Computational Linguistics ACL 2024*, 15747–15760. https://doi.org/10.18653/v1/2024.findings-acl.931

Parra-Gallego, L. F., & Orozco-Arroyave, J. R. (2022). Classification of emotions and evaluation of customer satisfaction from speech in real world acoustic environments. *Digital Signal Processing*, *120*, 103286. https://doi.org/10.1016/j.dsp.2021.103286

Pepino, L., Riera, P., & Ferrer, L. (2021). Emotion recognition from speech using wav2vec 2.0 embeddings. *Interspeech 2021*, 3400–3404. https://doi.org/10.21437/Interspeech.2021-703

Sutejo, B. S. I., Aryani, D., & Asianto, A. (2025). Pemodelan pengaruh risiko dengan pendekatan centrality dan graph-based learning approach pada transformasi digital kesehatan. *JUKI : Jurnal Komputer Dan Informatika*, *7*(2), 249–258. https://doi.org/10.53842/juki.v7i2.1992

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312–323. https://doi.org/10.1016/j.bspc.2018.08.035