Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) p-ISSN: 2355-7699
Vol. 13, No. 1, Februari 2026, hlm. 189-198 e-ISSN: 2528-6579

EVALUASI DAN OPTIMALISASI MODEL CNN-*TRANSFORMER ENCODER* DALAM DETEKSI STRES MELALUI SINYAL SUARA

Barlian Henryranu Prasetio[*1], Edita Rosana Widasari[2], Syifa' Hukma Shabiyya[3]

[1,2]Fakultas Ilmu Komputer, Universitas Brawijaya, Malang, [3]Magister Ilmu Komputer, Universitas Brawijaya, Malang
Email: [1]barlian@ub.ac.id, [2]editarosanaw@ub.ac.id, [3]syifahukma_s@student.ub.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 13 September 2024, diterima untuk diterbitkan: 23 Oktober 2025)

Abstrak

Deteksi stres melalui sinyal suara masih menghadapi tantangan akurasi karena keterbatasan model konvensional dalam menangkap distribusi frekuensi spasial-temporal. Oleh karena itu, diperlukan pendekatan baru yang mampu mengekstraksi pola kompleks secara efektif. Artikel ini mengeksplorasi peningkatan performa deteksi stres melalui sinyal suara dengan mengintegrasikan model *Convolutional Neural Network* (CNN) dan *Transformer Encoder*. Kami mengevaluasi berbagai konfigurasi jumlah *head* pada *self-attention* dan nilai *learning rate* untuk model *CNN-Transformer Encoder* guna mengidentifikasi parameter optimal. Hasil eksperimen menunjukkan bahwa konfigurasi dengan 6 *head* pada *Transformer Encoder* dan *learning rate* 0,01 memberikan performa terbaik dengan nilai loss terendah sebesar 0,5034, akurasi tertinggi 78,37%, serta peningkatan pada *precision, recall,* dan *F1-score*. Selain itu, penggabungan model CNN dengan *Transformer Encoder* secara paralel secara signifikan meningkatkan akurasi deteksi stres dibandingkan dengan model *baseline* CNN dan DSCNN. Pengujian lebih lanjut menggunakan *confusion matrix* menunjukkan keunggulan model DSCNN-*Transformer Encoder* dalam mendeteksi kelas stres dengan akurasi tertinggi. Pengujian pada dataset yang berbeda juga menunjukkan bahwa model yang diusulkan memiliki kestabilan yang baik. Temuan ini menegaskan efektivitas integrasi *Transformer Encoder* dalam meningkatkan performa deteksi stres pada sinyal suara.

Kata kunci: *Deteksi Stres, Convolutional Neural Network (CNN), Transformer Encoder, Learning Rate, Self-Attention, Sinyal Suara*

EVALUATION AND OPTIMIZATION OF CNN-TRANSFORMER ENCODER MODEL FOR STRESS DETECTION THROUGH SPEECH SIGNALS

Abstract

Stress detection through speech signals still faces accuracy challenges due to the limitations of conventional models in capturing spatial-temporal frequency distributions. Therefore, new approaches are needed that can effectively extract complex patterns. This study explores enhancing stress detection performance through speech signals by integrating Convolutional Neural Network (CNN) and Transformer Encoder models. We evaluated various configurations of self-attention head counts and learning rates for the CNN-Transformer Encoder model to identify optimal parameters. Experimental results indicate that a configuration with 6 heads in the Transformer Encoder and a learning rate of 0.01 yields the best performance with the lowest loss of 0.5034, highest accuracy of 78.37%, and improvements in precision, recall, and F1-score. Furthermore, the parallel integration of CNN with Transformer Encoder significantly improves stress detection accuracy compared to baseline CNN and DSCNN models. Further analysis using confusion matrices highlights the superior performance of the DSCNN-Transformer Encoder model in detecting stress classes with the highest accuracy. These findings affirm the effectiveness of integrating Transformer Encoder in enhancing stress detection performance from voice signals.

Keywords: *Stress Detection, Convolutional Neural Network (CNN), Transformer Encoder, Learning Rate, Self-Attention, Speech Signal*

1. PENDAHULUAN

Stres merupakan respons fisiologis terhadap tekanan mental, emosional, atau fisik (Useche et al., 2017). Berdasarkan survei *Health Service Monitor* (Ipsos, 2023), stres menjadi masalah kesehatan ketiga yang paling dikhawatirkan oleh responden, dengan angka mencapai 30%. Faktor penyebab stres meliputi

tuntutan tugas psikologis yang tinggi, kurangnya kesempatan berkembang, aspek sosial negatif, dan aspek organisasi negatif (Harmsen et al., 2018). Kegagalan beradaptasi dengan stres dapat menyebabkan malfungsi otak, masalah fisiologis, serta tantangan psikologis seperti depresi, kecemasan, rasa sakit, kelelahan, dan berbagai gejala lain (Gulzhaina et al., 2018).

Pengukuran stres sering dilakukan menggunakan sensor fisik seperti sensor *photoplethysmography* (PPG) untuk melacak aktivitas jantung atau mengukur kadar hormon stres dalam darah (Yun et al., 2022), serta pengukuran berbasis gambar wajah (Jeon et al., 2021). Namun, metode ini sering menimbulkan ketidaknyamanan dan rasa canggung pada individu. Han et al. (2018) menyarankan penggunaan pengukuran stres melalui suara, yang dianggap lebih mudah, non-invasif, dan dapat dilakukan dengan mikrofon tersembunyi tanpa perlu menempelkan alat pada tubuh penderita.

Speech Stress Recognition (SSR) adalah otomatisasi untuk mengidentifikasi tingkat stres melalui analisis suara. Stres menyebabkan ketegangan otot di tubuh, termasuk pita suara, yang mengubah karakteristik suara (Folk, 2021). Ketegangan otot dan laju pernapasan yang meningkat saat stres mengubah mekanika produksi suara dan mempengaruhi frekuensi suara yang dihasilkan (Slavich et al., 2019). Tingkat stres dapat dideteksi melalui perubahan frekuensi ini. Metode ekstraksi fitur sinyal suara yang umum digunakan adalah *Mel-Frequency Cepstral Coefficients* (MFCC), yang meniru cara kerja pendengaran manusia dan terbukti efektif dalam mendeteksi stres (Hilmy et al., 2021; Abdul et al, 2022).

Penerapan pembelajaran mesin dalam pengenalan suara terus berkembang signifikan. *Support Vector Machine* (SVM) adalah salah satu algoritme yang menjanjikan untuk berbagai tugas seperti deteksi peradangan organ (Chui et al, 2019) dan pengenalan emosi suara (Sun et al., 2019). Namun, SVM memiliki keterbatasan dalam menangani data tidak terstruktur seperti suara dan ketergantungan pada metode lain untuk klasifikasi multi-class (Kamil et al., 2016).

Convolutional Neural Network (CNN), yang diperkenalkan oleh LeCun pada tahun 1989, menjadi populer karena kemampuannya dalam tugas klasifikasi suara multi-class (Massoudi et al., 2021). Penelitian membandingkan performa SVM dan CNN untuk klasifikasi suara stres menunjukkan bahwa CNN mencapai akurasi terbaik (Shahin et al., 2021). Namun, CNN memiliki kekurangan dalam menangkap distribusi frekuensi sinyal suara stres yang tersebar dalam rentang waktu (Sun et al., 2022).

Kekurangan tersebut dapat diatasi dengan melakukan modifikasi pada CNN. Salah satu pendekatan adalah menggabungkan CNN dengan *Recurrent Neural Network* (RNN) secara serial (Choi et al., 2017). Namun, model serial memiliki kelemahan dalam menangani data emosional yang tersebar dan tidak berpola. Jiang et al. (2019) mengajukan model *Parallelized Convolutional Recurrent Neural Network* (PCRN) untuk meningkatkan pemahaman perubahan emosi, tetapi masih memiliki keterbatasan dalam mengidentifikasi informasi relevan dari data yang tersebar dan tidak teratur.

Pada tahun 2017, Vaswani et al. memperkenalkan algoritme *Transformer* yang mampu menangkap informasi fitur spasial dan memprediksi distribusi frekuensi yang tersebar pada suara stres dalam rentang waktu tertentu (Bautista et al., 2022). *Transformer* menggunakan mekanisme *self-attention* yang menghubungkan blok-bloknya untuk mempelajari hubungan antara fitur-fitur suara pada berbagai rentang waktu, efektif dalam mengenali ucapan stres (Al-onazi et al., 2022).

Dengan meningkatnya kebutuhan sistem monitoring kesehatan mental yang non-invasif, sistem deteksi stres berbasis suara yang akurat menjadi krusial untuk aplikasi di bidang kesehatan digital, manajemen SDM, dan keselamatan transportasi. Oleh karena itu, integrasi CNN dan Transformer Encoder diharapkan memberikan solusi yang lebih andal dalam mendeteksi stres secara real-time.

Penelitian ini bertujuan meningkatkan identifikasi level stres berdasarkan ucapan berbasis ekstraksi fitur MFCC, kemudian CNN akan mewakili fitur-fitur spasial, sementara *Transformer Encoder* akan mengatasi kekurangan CNN dalam menangkap distribusi frekuensi suara stres, sehingga meningkatkan keakuratan pengenalan suara stres. Peningkatan performansi dan kinerja algoritme akan diteliti melalui pengujian dan perhitungan metrik evaluasi berupa *loss,* akurasi*, precision, recall*, dan *F1-score* dengan pengujian parameter jumlah *head*, *learning rate*, dan perbandingan performa antara *baseline* dan *proposed method*. Kemudian, pengujian pada dataset yang berbeda juga dilakukan untuk mengetahui kestabilan model yang diusulkan.

2. KAJIAN PUSTAKA

2.1. Studi Terkait

Pengenalan informasi emosi dalam ucapan merupakan tantangan di bidang kecerdasan buatan. Jiang et al. (2019) mengusulkan *Parallelized Convolutional Recurrent Neural Network* (PCRN) dengan fitur spektral untuk pengenalan emosi dalam ucapan. Model ini menggabungkan CNN dan LSTM secara paralel untuk memproses dua jenis fitur yang berbeda secara bersamaan, sehingga mempelajari perubahan halus dalam emosi dengan lebih baik. Model PCRN menangkap perubahan emosional dalam domain waktu-frekuensi menggunakan database seperti CASIA, EMO-DB, ABC, dan SAVEE. Hasil akurasi untuk setiap database adalah: CASIA kurang dari 50%, EMO-DB kurang dari 80%,

ABC sekitar 47,62% hingga 55,70%, dan SAVEE kurang dari 70%.

Bezoui et al. (2017) menggunakan metode *Mel-Frequency Cepstral Coefficients* (MFCC) untuk ekstraksi fitur dalam pengenalan suara pada bacaan Al-Quran. Penelitian ini menguji jumlah filter MFCC dan tipe window yang digunakan. Jumlah filter diuji dengan variasi 12, 22, 32, dan 42, dengan hasil terbaik pada 32 filter, menghasilkan efisiensi 85%. Tipe window hamming menunjukkan performa lebih baik daripada rectangular dengan efisiensi 75% dibandingkan 55%.

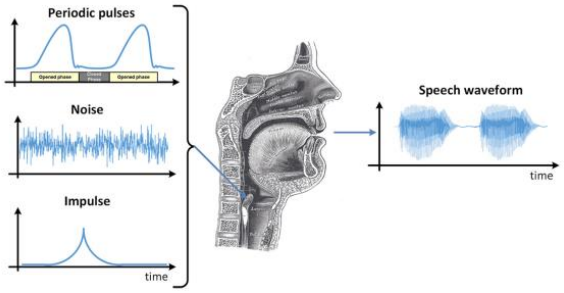
Vaswani et al. (2017) memperkenalkan metode *Transformer* dalam penelitian "*Attention Is All You Need*," yang mengusulkan arsitektur jaringan sederhana berdasarkan mekanisme attention tanpa pengulangan dan konvolusi. Model ini terdiri dari *encoder* dan *decoder*, masing-masing dengan enam blok identik. Setiap lapisan dalam *encoder* memiliki dua sub-lapisan: *multi-head self-attention mechanism* dan *fully connected feed-forward network.*

Mustaqeem et al. (2020) mengusulkan *Deep Stride CNN Architecture* (DSCNN) untuk deteksi emosi suara menggunakan dataset IEMOCAP dan RAVDESS. Metode ini menggunakan arsitektur CNN tanpa skema pooling pada layernya untuk mengekstraksi fitur-fitur tingkat tinggi dari spektogram sinyal ucapan dan mendeteksi pola tersembunyi dalam lapisan konvolusi.

Berbeda dengan model PCRN yang menggabungkan CNN dan LSTM secara paralel, studi ini mengusulkan integrasi CNN *dengan Transformer Encoder*. CNN menangkap fitur-fitur spasial, sedangkan *Transformer Encoder* mengatasi kekurangan CNN dalam menangkap distribusi frekuensi suara stres yang tersebar dalam rentang waktu. *Transformer* menggunakan mekanisme *self-attention* yang menghubungkan blok-bloknya, efektif dalam mengenali ucapan stres tanpa bergantung pada pola-pola berulang. Hal ini diharapkan dapat meningkatkan akurasi pengenalan suara stres secara signifikan, mengatasi keterbatasan model sebelumnya dalam menangani data yang tersebar dan tidak berpola.

2.2. Stres dan Ucapan

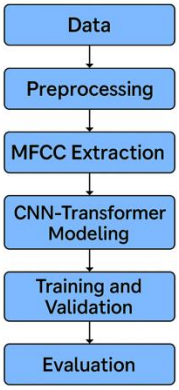
Ucapan adalah cara alami mengekspresikan diri, dan kini digunakan dalam aplikasi komputer (Akçay, 2020). Ucapan stres melibatkan gerakan kompleks alat artikulasi dan sistem pernapasan. Analisis stres suara meliputi aspek verbal dan non-verbal. Verbal adalah suara yang dikeluarkan saat berkomunikasi, sedangkan non-verbal mencakup kecepatan bicara, volume, intonasi, dan ketidakstabilan vokal, yang mencerminkan tingkat stres (Jaafar & Lachiri, 2022). Gambar 1 menunjukkan bahwa terdapat tiga komponen sinyal ucapan adalah sumber suara, saluran vokal, dan sinyal ucapan (output). Stres memengaruhi ketegangan otot pita suara, posisi artikulator, dan frekuensi suara (Giannakakis et al., 2022).



Gambar 1. Ilustrasi Mekanisme Produksi Suara Manusia

3. METODE PENELITIAN

Secara keseluruhan, tahapan penelitian ini mencakup proses dari pengumpulan data hingga evaluasi performa model, sebagaimana ditunjukkan pada Gambar 2.



Gambar 2. Diagram Alur Penelitian Deteksi Stres Berbasis CNN-Transformer Encoder

3.1. Dataset

Penelitian ini menggunakan data yang diperoleh dari *Speech Under Simulated and Actual Stress* (*SUSAS*) yang dibuat dibawah arahan Prof. John H. L. Hansen (Hansen, 1999) yang disponsori oleh *Air Force Research Laboratory*. Data terbagi menjadi empat domain dengan berbagai emosi dan tekanan. Terdapat 32 partisipan yang terdiri dari 19 orang laki-laki dan 13 orang perempuan dengan rentang usia 22-76 tahun yang menghasilkan lebih dari 16.000 ucapan. Data-data ini diambil dari mikrofon dengan bahasa Inggris yang dapat diaplikasikan untuk *speech recognition*. Terdapat 35 kata yang membentuk database *SUSAS* yang diambil menggunakan konverter A/D 16-bit dengan kecepatan sampel 8kHz.

Untuk menguji performa model, penelitian ini mengambil sebanyak 2807 sampel dari kumpulan data *SUSAS*. Data diperoleh pada link https://catalog.ldc.upenn.edu/LDC99S78. Data ini diambil dari tujuh pembicara, termasuk tiga perempuan dan empat laki-laki. Data dikategorikan ke dalam lima kelas: "*Angry*", "*High stress*", "*Low stress*", "*Neutral*", and "*Soft*". Jumlah data yang digunakan ditunjukkan pada Tabel 2.

Tabel 2. Dataset SUSAS yang Digunakan

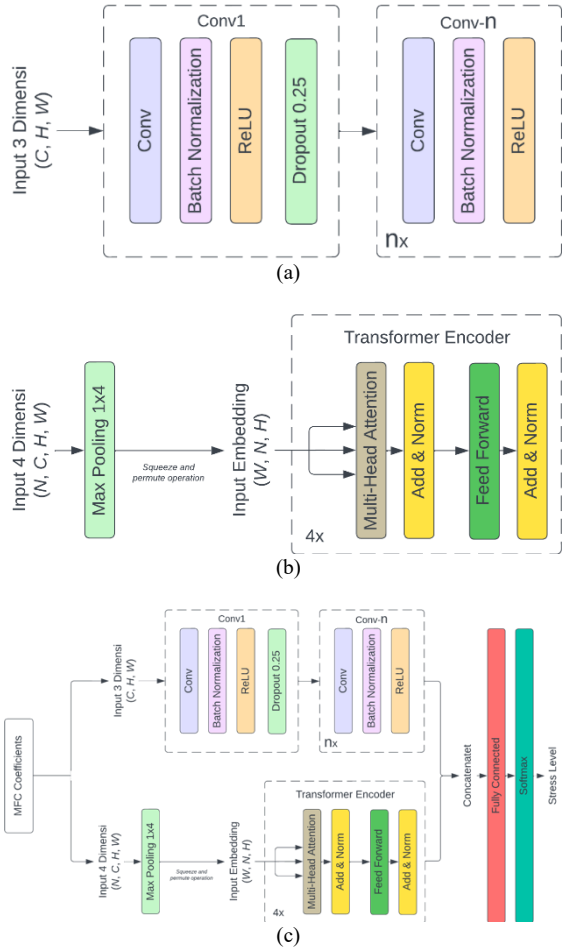
Kelas Stres	Jenis Kelamin	
	Laki-laki	Perempuan
Angry	210	226
High stress	319	301
Low stress	319	301
Neutral	319	301
Soft	210	301

Selain menggunakan dataset SUSAS, penelitian ini juga memanfaatkan dataset *Ryerson Audio-Visual Database of Emotional Speech and Song* (RAVDESS) guna menguji kestabilan model yang diusulkan (Livingstone et al, 2018). RAVDESS adalah dataset yang berisi 7356 file yang melibatkan 24 subjek, terdiri dari 12 laki-laki dan 12 perempuan. Dataset ini mencakup 8 emosi: "*calm*", "*happy*", "*sad*", "*angry*", "*fearful*", "*disgust*", "*surprised*", dan "*neutral*", dan tersedia dalam format audio dan visual. Untuk penelitian ini, hanya file audio yang digunakan. Data suara ini direkam dengan kecepatan sampel 48 kHz menggunakan mikrofon berkualitas tinggi, dan kategori emosi yang ada dipetakan sesuai dengan kategori stres yang dianalisis dalam dataset SUSAS. Pengujian performa model pada dataset RAVDESS bertujuan untuk mengevaluasi kestabilan model CNN-*Transformer Encoder* dalam mendeteksi stres pada kondisi dan dataset yang berbeda.

Setiap file audio melalui proses normalisasi dan framing menggunakan window Hamming 25 ms dengan 10 ms overlap. Dari hasil framing, diekstraksi 13 koefisien MFCC per frame menggunakan 40 filter Mel, menghasilkan matriks fitur berukuran 40×13 sebagai input ke CNN-Transformer.

3.2 Arsitektur Sistem yang diusulkan

Penelitian ini menggunakan model paralelisasi CNN dengan *Transformer Encoder*. Kedua model tersebut menerima input dari hasil ekstraksi fitur MFCC dan menyesuaikannya dengan ukuran tensor yang dibutuhkan. Output *embedding* yang dihasilkan oleh model CNN (Gambar 3(a)) dan *Transformer Encoder* (Gambar 3(b)) akan di-concatenate menjadi *complete embedding* dan diteruskan ke *fully connected* (fc) layer. Model ini memiliki satu fc layer dengan 512 neuron yang mengeluarkan sejumlah kelas, kemudian diteruskan ke fungsi softmax untuk prediksi kelas stres. Ilustrasi arsitektur CNN-*Transformer Encoder* ditunjukkan pada Gambar 3(c). Studi ini mengusulkan untuk menggabungkan kekuatan CNN dalam menangkap fitur spasial dan *Transformer Encoder* dalam menangani distribusi frekuensi suara stres yang tersebar dalam rentang waktu, sehingga diharapkan mampu meningkatkan akurasi dan kinerja dalam mendeteksi stres melalui sinyal suara.



Gambar 3. Arsitektur Metode yang Diusulkan: a) Layer yang Digunakan pada Blok *Deep Stride Convolutional Neural Network*, b) Arsitektur *Transformer Encoder*, c) Arsitektur lengkap *Metode yang diusulkan*

Gambar 3(a) menggambarkan layer yang digunakan dalam blok *Deep Stride Convolutional Neural Network* (DSCNN). Input yang diterima oleh blok DSCNN dalam penelitian ini memiliki ukuran tiga dimensi: *Channel* (C), *Height* (H), dan *Width* (W). *Channel* (C) merupakan input dari hasil MFCC, Height (H) adalah dimensi frekuensi dari fitur input *spectrogram* atau MFCC, dan *Width* (W) adalah dimensi waktu dari fitur input spectrogram atau MFCC. Penelitian ini menggunakan tujuh blok, masing-masing terdiri dari *convolutional layer*, 2x2 *stride*, 2x2 *padding*, diikuti oleh *batch normalization layer* dan fungsi aktivasi ReLU. Blok konvolusi pertama memiliki tambahan *layer dropout*. Rincian blok pertama memiliki *filter* ukuran 16 dengan *kernel* 7x7 dan *dropout* 0,25. Blok kedua dan ketiga memiliki filter 32 dengan *kernel* 5x5 dan 3x3. Blok keempat dan kelima memiliki filter 64 dengan *kernel* 3x3. Blok keenam dan ketujuh memiliki *filter* 128 dengan *kernel* 3x3. Output *embedding* dari blok CNN terakhir dikonversi menjadi vektor melalui *flattening layer*. Parameter detail untuk DSCNN ditunjukkan pada Tabel 1. Pendekatan ini memaksimalkan ekstraksi fitur dari input MFCC, memastikan

informasi yang relevan dipertahankan untuk analisis lebih lanjut.

Tabel 1. Parameter Model *Baseline* DSCNN

Tipe	Parameter
Conv2d	in_channels=1, out_channels=32, kernel_size=7, stride=2, padding=2
BatchNorm2d	num_features=32
ReLU	
Dropout	p=0,25
Conv2d	in_channels=32, out_channels=32, kernel_size=5, stride=2, padding=2
BatchNorm2d	num_features=32
ReLU	
Conv2d	in_channels=32, out_channels=64, kernel_size=3, stride=2, padding=2
BatchNorm2d	num_features=64
ReLU	
Conv2d	in_channels=64, out_channels=64, kernel_size=3, stride=2, padding=2
BatchNorm2d	num_features=64
ReLU	
Conv2d	in_channels=64, out_channels=128, kernel_size=3, stride=2, padding=2
BatchNorm2d	num_features=128
ReLU	
Conv2d	in_channels=128, out_channels=128, kernel_size=3, stride=2, padding=2
BatchNorm2d	num_features=128
ReLU	
Conv2d	in_channels=128, out_channels=512, kernel_size=3, stride=2, padding=2
BatchNorm2d	num_features=512
ReLU	

Gambar 2(b) menunjukkan input (fitur MFCC) melewati lapisan *max pooling* 1x4 untuk mengurangi dimensi waktu. Format tensor awal adalah N, C, H, W. Operasi squeeze kemudian digunakan untuk menghapus dimensi C, menghasilkan tensor dengan urutan N, H, dan W. Untuk mencocokkan urutan input yang diperlukan oleh *Transformer Encoder*, tensor diubah menggunakan metode *permute* dari N, H, W menjadi W, N, H. Setelah penyesuaian, tensor dimasukkan ke dalam *Transformer Encoder*, yang terdiri dari 4 blok identik. Output dari *Transformer Encoder* digabungkan dengan output model CNN dan diteruskan ke *fully connected layer*.

4. HASIL DAN PEMBAHASAN

Untuk mengevaluasi keefektifan sistem yang diusulkan, sistem dievaluasi dan analisis kinerja sistem yang diusulkan dalam beberapa indikator yaitu: evaluasi jumlah *head* pada *self-attention* di *Transformer Encoder* terhadap performa model, evaluasi *learning rate* terhadap kinerja model CNN-*Transformer Encoder*, dan evaluasi perbandingan kinerja metode yang diusulkan dengan model *baseline*.

Pada tahap pelatihan, *optimizer* yang digunakan yaitu *Adam* dengan *weight decay* sebesar $1e^{-3}$. Jumlah *epoch* yang digunakan adalah 200 dengan jumlah *batch* adalah 64.

4.1. Evaluasi Jumlah Head

Pengujian pertama mengevaluasi pengaruh jumlah *head* pada *self-attention* dalam arsitektur *Transformer Encoder* terhadap deteksi stres dari sinyal suara, dengan jumlah head yang diuji adalah 1, 2, 4, dan 6 (Shin et al., 2020). Pengujian dilakukan pada model DSCNN-*Transformer Encoder*.

Tabel 3. Hasil Pengujian Jumlah *Head* pada *Self-Attention* di *Transformer Encoder* Terhadap Performa Model

Jumlah *Head*	*Loss*	Akurasi (%)	*Precision*	*Recall*	*F1-Score*
1	0,6696	75,5319	0,7824	0,7739	0,7717
2	0,6715	75,5319	0,7703	0,7683	0,7690
4	0,6071	77,6596	0,7901	0,7810	0,7809
6	0,5034	78,3688	0,8073	0,8004	0,7980

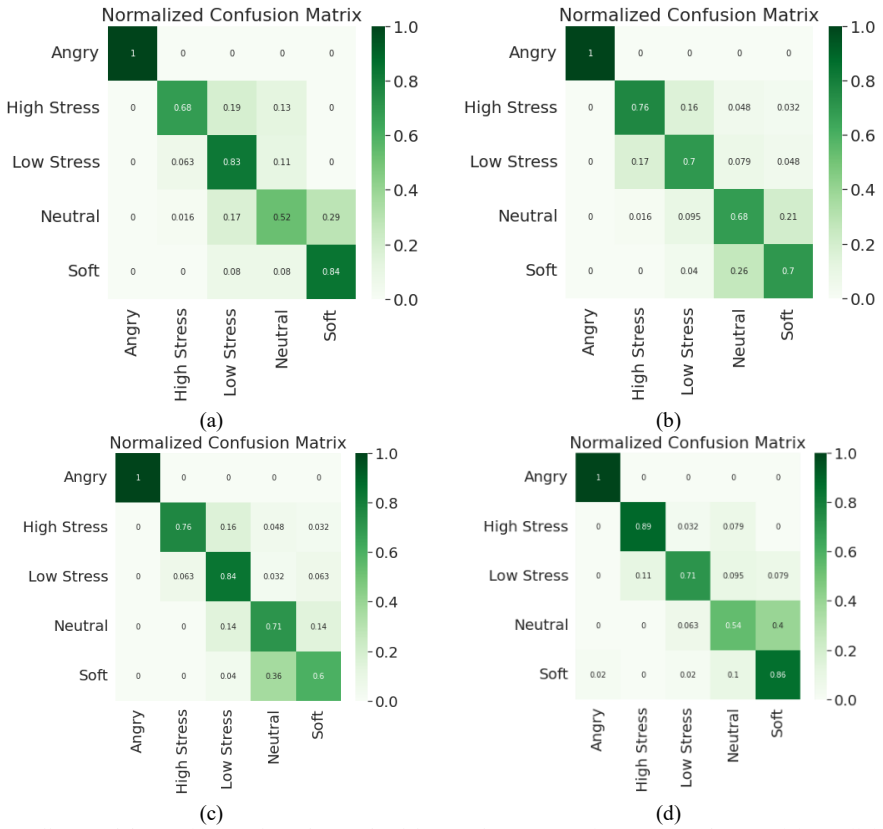
Hasil menunjukkan (Tabel 3) bahwa jumlah *head* mempengaruhi metrik evaluasi secara signifikan. Dengan 1 *head*, nilai *loss* adalah 0,6696, akurasi 75,5319%, serta *precision, recall*, dan *F1-score* masing-masing 0,7824, 0,7739, dan 0,7717. Pada 2 *head, loss* sedikit meningkat menjadi 0,6715, namun akurasi tetap sama. Dengan 4 head, terjadi penurunan *loss* menjadi 0,6071 dan akurasi meningkat menjadi 77,6596%, dengan *precision, recall*, dan *F1-score* yang juga meningkat. Jumlah *head* 6 mencapai kinerja terbaik dengan *loss* terendah 0,5034, akurasi tertinggi 78,3688%, serta *precision, recall*, dan *F1-score* terbaik.

Analisis *confusion matrix* (Gambar 4) menunjukkan prediksi yang sangat baik pada kelas *Angry* untuk semua jumlah *head*. Untuk kelas *High Stress*, *head* 6 memberikan hasil terbaik (0,89), Kelas *Low Stress* memberikan hasil terbaik dengan *head* 4 (0,84), sedangkan kelas *Soft* juga menunjukkan hasil terbaik dengan *head* 6 (0,86). Secara keseluruhan, *head* 6 adalah yang paling optimal.

4.2. Evaluasi Learning Rate

Pengujian kedua mengevaluasi pengaruh *learning rate* (lr) terhadap metrik evaluasi model CNN-*Transformer Encoder*, dengan nilai lr yang diuji adalah 0,01, 0,005, dan 0,001. *Learning rate* yang tepat krusial untuk performa model yang optimal, menghindari *overfitting* atau *underfitting*. Pengujian dilakukan dengan optimasi Adam dan epoch sebanyak 200.

Pada Tabel 3, untuk lr 0,001, model menunjukkan nilai loss sebesar 0,6785 dan akurasi 73,0496%, dengan *precision, recall*, dan *F1-score* yang baik tetapi tidak optimal. Dengan lr 0,005, terdapat peningkatan signifikan, yaitu *loss* menurun menjadi 0,6727 dan akurasi meningkat menjadi 76,5957%, serta peningkatan pada *precision, recall,* dan *F1-score*. Namun, lr 0,01 menghasilkan performa terbaik dengan *loss* 0,5034 dan akurasi 78,3688%, meskipun perbedaan pada *precision* dan *recall* dibandingkan lr 0,005 cukup kecil.



Gambar 4. Hasil Pengujian Jumlah *Head* yang Digunakan pada Sisi *Transformer Encoder* Menggunakan *Model Parallel CNN-Transformer Encoder*: (a) *Confusion Matrix* dengan Jumlah *Head* 1, (b) *Confusion Matrix* dengan Jumlah *Head* 2, (c) *Confusion Matrix* dengan Jumlah *Head* 4, dan (d) *Confusion Matrix* dengan Jumlah *Head* 6

Tabel 3. Hasil Pengujian *Learning Rate* Terhadap Performa Model CNN-*Transformer Encoder*

lr	Loss	Akurasi (%)	Precision	Recall	F1-Score
0,001	0,6785	73,0496	0,7492	0,7361	0,7325
0,005	0,6727	76,5957	0,8110	0,7882	0,7789
0,01	0,5034	78,3688	0,8073	0,8004	0.7980

Confusion matrix (Gambar 5) mengungkapkan prediksi kelas yang sangat baik pada kelas *Angry* untuk semua lr. Pada kelas *High Stress*, lr 0,01 menghasilkan prediksi terbaik (0,89), sedangkan lr 0,001 menghasilkan nilai tertinggi pada kelas *Low Stress* (0,89), Kelas Neutral memiliki prediksi terbaik pada lr 0,001 (0,6), dan kelas *Soft* terbaik pada lr 0,005 (0,94). Pemilihan lr yang tepat sangat penting untuk meningkatkan kualitas prediksi model.

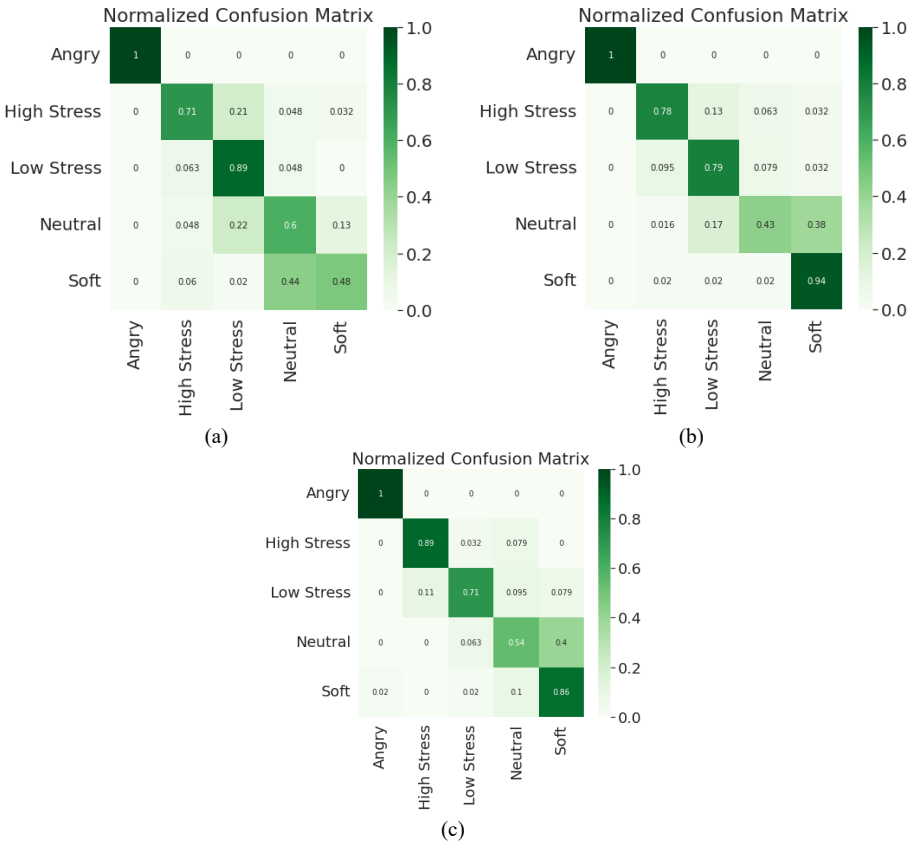
4.3. Perbandingan Kinerja

Empat model *baseline* dibandingkan dengan model Transformer Encoder untuk meningkatkan deteksi stres melalui sinyal suara. Keempat model dilatih dan diuji menggunakan data yang sama untuk mengevaluasi performa mereka. Hasil pengujian menunjukkan (Tabel 4) bahwa model *baseline* CNN-Bi LSTM memiliki akurasi 71,9858% dan nilai *loss* 0,7656, dengan *precision, recall,* dan *F1-score* masing-masing 0,6980, 0,6700, dan 0,6666. Model kedua, CNN *Gated Recurrent Unit* (GRU), menunjukkan peningkatan performa dengan akurasi 71,6312% dan nilai *loss* 0,8467, serta *precision* 0,6900, *recall* 0,6980, dan *F1-score* 0,6880.

Model *baseline* ketiga CNN (Bautista et al., 2022) memiliki akurasi 69,86% dan nilai *loss* 0,7420, dengan *precision, recall,* dan *F1-score* masing-masing 0,7328, 0,7060, dan 0,7055. Model keempat, DSCNN (Mustaqeem & Kwon, 2020), menunjukkan peningkatan performa dengan akurasi 75,8865% dan nilai *loss* 0,6126, serta *precision* 0,7752, *recall* 0,7432, dan *F1-score* 0,7454.

Model kelima, CNN-*Transformer Encoder*, memperoleh akurasi 73,76% dengan loss 0,6550, serta *precision* 0,7627, *recall* 0,7526, dan *F1-score* 0,7547. Model keempat, DSCNN-*Transformer Encoder*, menunjukkan performa terbaik dengan akurasi 78,3688%, nilai *loss* 0,5034, *precision* 0,8073, *recall* 0,8004, dan *F1-score* 0,7980. Integrasi *Transformer Encoder* dengan model *baseline* CNN atau DSCNN secara paralel terbukti efektif dalam meningkatkan deteksi stres.

Confusion matrix menunjukkan bahwa model *DSCNN-Transformer Encoder* menghasilkan prediksi yang paling akurat untuk kelas *High Stress* (0,89) dan *Soft* (0,86), sedangkan model CNN menunjukkan hasil yang lebih rendah untuk kelas-kelas tertentu. Hasil ini menunjukkan keunggulan model yang mengintegrasikan *Transformer Encoder* dalam mendeteksi stres dari sinyal suara (Gambar 6).



Gambar 5. Hasil Pengujian Nilai *Learning Rate* yang Digunakan Menggunakan *Model Parallel* CNN-*Transformer Encoder* (a) *Confusion Matrix* dengan *Learning Rate* 0,001, (b) *Confusion Matrix* dengan *Learning Rate* 0,005, dan (c) *Confusion Matrix* dengan *Learning Rate* 0,01

Tabel 4. Hasil Pengujian Perbandingan Metrik Evaluasi antara Model *Baseline* dengan Proposed Method

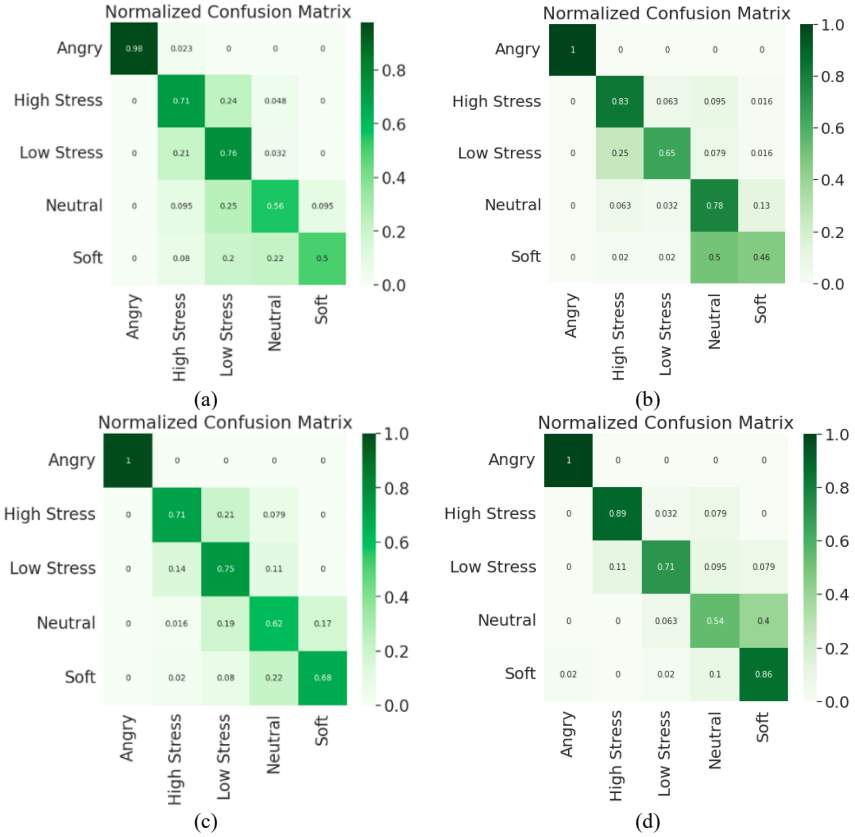
Model	Loss	Akurasi (%)	Precision	Recall	F1-Score
CNN-Bi LSTM	0,7656	71,9858	0,6980	0,6700	0,6666
CNN-GRU	0,8467	71,6312	0,6900	0,6980	0,6880
CNN (Bautista et al., 2022)	0,7420	69,8600	0,7328	0,7060	0,7055
DSCNN (Mustaqeem et al., 2020)	0,6126	75,8865	0,7752	0,7432	0,7454
CNN-Transformer Encoder (*proposed method*)	0,6550	73,7600	0,7627	0,7526	0,7547
DSCNN-Transformer Encoder (*proposed method*)	**0,5034**	**78,3688**	**0,8073**	**0,8004**	**0,7980**

Penelitian ini juga menguji dan membandingkan *proposed method* CNN-*Transformer Encoder* serta DSCNN-*Transformer Encoder* pada dua dataset, yaitu SUSAS dan RAVDESS untuk menunjukkan kestabilan model yang diusulkan. Penggunaan dua dataset yang memiliki karakteristik berbeda bertujuan untuk menilai kemampuan generalisasi model dalam mendeteksi stres dari sinyal suara di berbagai kondisi lingkungan dan emosi.

Hasil pengujian performa dari model yang diusulkan, pada kedua dataset SUSAS dan RAVDESS menunjukkan bahwa kedua model memiliki kestabilan yang baik. Seperti ditunjukkan pada Tabel 5, model CNN-*Transformer Encoder* pada dataset RAVDESS menghasilkan nilai loss sebesar 0,6893 dengan akurasi 74,13%, *precision* 0,7569, *recall* 0,7347, dan *F1-score* 0,7241. Sementara itu, pada dataset SUSAS, model yang sama menunjukkan *loss* sebesar 0,655 dengan akurasi yang sedikit lebih rendah, yaitu 73,76%, namun tetap menunjukkan hasil yang stabil dengan *precision* 0,7627, *recall* 0,7526, dan *F1-score* 0,7547. Meskipun terdapat sedikit perbedaan performa antara kedua dataset, kestabilan model ini tercermin dari konsistensi metrik evaluasi yang tetap mendekati hasil optimal.

Tabel 5. Hasil Pengujian Proposed Method Pada Dataset Berbeda

Dataset	Loss	Akurasi (%)	Precision	Recall	F1-Score
CNN-Transformer Encoder (*proposed method*)					
RAVDESS	0,6893	74,1259	0,7569	0,7347	0,7241
SUSAS	0,6550	73,760	0,7627	0,7526	0,7547
DSCNN-Transformer Encoder (*proposed method*)					
RAVDESS	0,6330	77,7500	0,8019	0,7813	0,7761
SUSAS	0,5034	78,3688	0,8073	0,8004	0,798



Gambar 6. Hasil Pengujian Perbandingan Metrik Evaluasi antara Model *Baseline* dengan *Proposed Method* (a) *Confusion Matrix Model Baseline CNN*, (b) *Confusion Matrix Model Baseline DSCNN*, (c) *Confusion Matrix Model Proposed Method CNN-Transformer Encoder*, dan (d) *Confusion Matrix Model Proposed Method DSCNN-Transformer Encoder*

5. KESIMPULAN

Penelitian ini telah menunjukkan bahwa integrasi model CNN dan *Transformer Encoder* secara paralel dapat secara signifikan meningkatkan deteksi stres melalui sinyal suara.

Berdasarkan hasil pengujian, beberapa kesimpulan dari penelitian ini adalah sebagai berikut. Pengujian terhadap jumlah *head* pada *self-attention* dalam arsitektur *Transformer Encoder* menunjukkan bahwa jumlah *head* yang optimal adalah 6. Model dengan 6 head memperoleh *loss* 0,5034, akurasi 78,37%, *precision* 0,8073, *recall* 0,8004, dan *F1-score* 0,7980. Ini menunjukkan bahwa dengan jumlah *head* yang tepat, model CNN-*Transformer Encoder* dapat menangkap pola kompleks dalam data suara, sehingga meningkatkan deteksi stres. Selain itu, pengujian nilai *learning rate* mengungkapkan bahwa nilai optimal adalah 0,01, dengan hasil *loss* 0,5034, akurasi 78,37%, *precision* 0,8073, *recall* 0,8004, dan *F1-score* 0,7980. Nilai *learning rate* yang lebih kecil atau lebih besar tidak memberikan kinerja yang sama baiknya, menegaskan pentingnya memilih learning rate yang tepat.

Perbandingan antara model *baseline* dan model yang diajukan menunjukkan bahwa model CNN-*Transformer Encoder* dan DSCNN-*Transformer Encoder* memberikan peningkatan signifikan dalam deteksi stres. Model DSCNN-Transformer Encoder, khususnya, mencapai nilai loss terendah 0,5034 dan akurasi tertinggi 78,3688%, menunjukkan bahwa integrasi *Transformer Encoder* dengan CNN atau DSCNN secara paralel efektif dalam meningkatkan deteksi stres. Selanjutnya, kestabilan juga ditunjukan oleh *proposed method* dimana tidak hanya efektif dalam mendeteksi stres pada satu jenis dataset, tetapi juga dapat beradaptasi dengan baik pada dataset lain yang memiliki variasi dalam emosi stres, seperti yang ditunjukkan pada pengujian dengan dataset RAVDESS dan SUSAS.

Untuk penelitian selanjutnya, disarankan untuk menggunakan teknik augmentasi data seperti *pitch shifting, adding noise,* atau *time-stretching* guna meningkatkan variasi kondisi stres dan melakukan *tuning hyperparameter* lebih lanjut. Selain itu, menguji model pada dataset lain atau data *primer* penting untuk memastikan generalisasi model. Menggabungkan *Transformer Encoder* dengan model lain selain CNN juga bermanfaat untuk perbandingan performa. Terakhir, mengimplementasikan model ke dalam aplikasi atau *platform* yang fleksibel dapat memudahkan penggunaan oleh berbagai kalangan.

DAFTAR PUSTAKA

ABDUL, Z. K & AL-TABALANI, A. K. 2022. Mel Frequency Cepstral Coefficient and its applications: A Review. IEEE Access, 10, 122136-122158. https://doi.org/10.1109/ACCESS.2022.3223444

AKÇAY, M. B., & OĞUZ, K. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. In Speech Communication, 116, 56–76. Elsevier B.V. https://doi.org/10.1016/j.specom.2019.12.001

AL-ONAZI, B. B., NAUMAN, M. A., JAHANGIR, R., MALIK, M. M., ALKHAMMASH, E. H., & ELSHEWEY, A. M. 2022. Transformer-Based Multilingual Speech Emotion Recognition Using Data Augmentation and Feature Fusion. Applied Sciences (Switzerland), 12(18). https://doi.org/10.3390/app12189188

BAUTISTA, J. L., LEE, Y. K., & SHIN, H. S. 2022. Speech Emotion Recognition Based on Parallel CNN-Attention Networks with Multi-Fold Data Augmentation. Electronics (Switzerland), 11(23). https://doi.org/10.3390/electronics11233935

BEZOUI, M., ELMOUTAOUAKKIL, A., & BENI-HSSANE, A. 2017. Feature extraction of some Quranic recitation using Mel-Frequency Cepstral Coeficients (MFCC). International Conference on Multimedia Computing and Systems-Proceedings, 0, 127–131. https://doi.org/10.1109/ICMCS.2016.7905619

CHOI, K., FAZEKAS, G., MARK, S., & KYUNGHYUN, C. 2017. Convolutional recurrent neural networks for music classification. IEEE International Conference on Acoustics, Speech and Signal Processing, 6567.

CHUI, K. T., & LYTRAS, M. D. 2019. A novel MOGA-SVM multinomial classification for organ inflammation detection. Applied Sciences (Switzerland), 9(11). https://doi.org/10.3390/app9112284

FOLK, J. (2021, May 19). Voice Changes Anxiety Symptoms. Anxietycentre.Com. https://www.anxietycentre.com/anxiety-disorders/symptoms/voice-changes/#:~:text=Stress%20can%20affect%20the%20quality,our%20vocal%20quality%20and%20performance.

GIANNAKAKIS, G., GRIGORIADIS, D., GIANNAKAKI, K., SIMANTIRAKI, O., RONIOTIS, A., & TSIKNAKIS, M. 2022. Review on Psychological Stress Detection Using Biosignals. IEEE Transactions on Affective Computing, 13(1), 440–460. https://doi.org/10.1109/TAFFC.2019.2927337

GULZHAINA, K. K., AIGERIM, K. N., OSPAN, S. S., HANS, S. J., AUSTRIA, N. B. C., & COX, R. 2018. Stress management techniques for students. Proceedings of the International Conference on the Theory and Practice of Personality Formation in Modern Society (ICTPPFMS), 47–56. https://doi.org/10.2991/ictppfms-18.2018.10

HILMY, M. S. H., ASNAWI, A. L., JUSOH, A. Z., ABDULLAH, K., IBRAHIM, S. N., MOHD RAMLI, H. A., & MOHAMED AZMIN, N. F. 2021. Stress Classification based on Speech Analysis of MFCC Feature via Machine Learning. Proceedings of the 8th International Conference on Computer and Communication Engineering, ICCCE 2021, 339–343. https://doi.org/10.1109/ICCCE50029.2021.9467176

HAN, H., BYUN, K., & KANG, H. G. 2018. A deep learning-based stress detection algorithm with speech signal. AVSU 2018 - Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia, Co-Located with MM 2018, 11–15. https://doi.org/10.1145/3264869.3264875

HANSEN, J. H. L. 1999. SUSAS LDC99S78. Philadelphia: Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC99S78

HARMSEN, R., HELMS-LORENZ, M., MAULANA, R., & VAN VEEN, K. 2018. The relationship between beginning teachers' stress causes, stress responses, teaching behaviour and attrition. Teachers and Teaching: Theory and Practice, 24(6), 626–643. https://doi.org/10.1080/13540602.2018.1465404

IPSOS. 2023. Ipsos Global Health Service Monitor - 2023. Diakses dari: [https://www.ipsos.com/sites/default/files/ct/news/documents/2023-09/Ipsos-Global-Health-Service-Monitor-2023-WEB.pdf].

JAAFAR, N., & LACHIRI, Z. 2022. Stress Recognition from Speech by Combining Image-based Deep Spectrum and Text-based Features. 2022 IEEE Information Technologies and Smart Industrial Systems, ITSIS 2022. https://doi.org/10.1109/ITSIS56166.2022.10118402

JEON, T., BAE, H. B., LEE, Y., JANG, S., & LEE, S. 2021. Deep-learning-based stress recognition with spatial-temporal facial information. Sensors, 21(22). https://doi.org/10.3390/s21227498

JIANG, P., FU, H., TAO, H., LEI, P., & ZHAO, L. 2019. Parallelized Convolutional Recurrent Neural Network with Spectral Features for Speech Emotion Recognition. IEEE Access, 7, 90368–90377. https://doi.org/10.1109/ACCESS.2019.2927384

KAMIL, A.-Z., XOCAYEV, A., & RUSTAMOV, S. 2016. Speech Recognition using Support Vector Machines. IEEE 10th International Conference on Application of Information and Communication Technologies (AICT). https://doi.org/10.1109/ICAICT.2016.7991664

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), pp. e0196391. https://doi.org 10.1371/journal.pone.0196391.

MASSOUDI, M., VERMA, S., & JAIN, R. 2021. Urban Sound Classification using CNN. Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021, 583–589. https://doi.org/10.1109/ICICT50816.2021.9358621

MUSTAQEEM, & KWON, S. 2020. A CNN-assisted enhanced audio signal processing for speech emotion recognition. Sensors (Switzerland), 20(1). https://doi.org/10.3390/s20010183

SHAHIN, I., NASSIF, A. B., & HINDAWI, N. 2021. Speaker identification in stressful talking environments based on convolutional neural network. International Journal of Speech Technology, 24(4), 1055–1066. https://doi.org/10.1007/s10772-021-09869-1

SHIN, H.-K., HAN, H., BYUN, K., & KANG, H.-G. 2020. Speaker-invariant Psychological Stress Detection Using Attention-based Network. IEEE International Conference on Big Data and Smart Computing (BigComp). https://ieeexplore.ieee.org/document/9306384

SLAVICH, G. M., TAYLOR, S., & PICARD, R. W. 2019. Stress measurement using speech: Recent advancements, validation issues, and ethical and privacy considerations. In The National Center for Biotechnology Information, 22(4), 408–413. Taylor and Francis Ltd. https://doi.org/10.1080/10253890.2019.1584180

SUN, J., WANG, X., ZHAO, K., HAO, S., & Wang, T. 2022. Multi-Channel EEG Emotion Recognition Based on Parallel Transformer and 3D-Convolutional Neural Network. Mathematics, 10(17). https://doi.org/10.3390/math10173131

SUN, L., FU, S., & WANG, F. 2019. Decision tree SVM model with Fisher feature selection for speech emotion recognition. Eurasip Journal on Audio, Speech, and Music Processing, 2019(1). https://doi.org/10.1186/s13636-018-0145-5

USECHE, S. A., ORTIZ, V. G., & CENDALES, B. E. 2017. Stress-related psychosocial factors at work, fatigue, and risky driving behavior in bus rapid transport (BRT) drivers. Accident Analysis and Prevention, 104, 106–114. https://doi.org/10.1016/j.aap.2017.04.023

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L., & POLOSUKHIN, I. 2017. Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS). http://arxiv.org/abs/1706.03762

YUN, M., HONG, S., YOO, S., KIM, J., PARK, S. M., & LEE, Y. 2022. Lightweight End-to-End Stress Recognition using Binarized CNN-LSTM Models. Proceeding - IEEE International Conference on Artificial Intelligence Circuits and Systems, AICAS 2022, 270–273. https://doi.org/10.1109/AICAS54282.2022.9869974