Laman web jurnal: https://ejournal.unama.ac.id/index.php/processor

Jurnal Processor

P-ISSN: 1907-6738 | E-ISSN: 2528-0082

Pengenalan Pola Depresi Berbasis Suara Menggunakan Ekstraksi Fitur Mel-Frequency Cepstral Coefficients

Wahju Tjahjo Saputro[1*], Abdul Fadlil[2], Murinto[3]

[1]Program Studi Informatika, Universitas Ahmad Dahlan, Yogyakarta 55191, Indonesia.
[1]Program Studi Teknologi Informasi, Universitas Muhammadiyah Purworejo, Purworejo 54111, Indonesia.
[2]Departemen Teknik Elektro, Universitas Ahmad Dahlan, Yogyakarta 55191, Indonesia.
[3]Departemen Informatika, Universitas Ahmad Dahlan, Yogyakarta 55191, Indonesia.
*Penulis Korespondensi, Email: 2436083021@webmail.uad.ac.id

Abstrak− Penyakit depresi menjadi pemasalahan penting saat ini, karena ada peningkatan secara global penderita depresi. Faktor depresi banyak dan kompleks, dapat menjangkau semua kalangan baik anak-anak hingga lansia. Tujuan penelitian ini untuk mengetahui pola depresi dan sehat berdasarkan ekstraksi fitur suara. Metode ekstraksi fitur yang digunakan Mel-Frequency Cepstral Coefficients (MFCC). Ketiga model yang digunakan untuk mengukur performa dan evaluasi yaitu Naïve Bayes (NB), Decision Tree (DT), dan Random Forest (RD). Dataset yang digunakan EATD-Corpus berisi 162 rekaman mahasiswa Universitas Tongji Tiongkok. Hasil penelitian menunjukkan pola depresi dan sehat berhasil nampak dengan parameter MFCC yaitu 25 ukuran masing-masing frame, 10 jarak antar frame, alpha 0,97 sebagai nilai koefisien pre-emphasis, 40 jumlah maksimum koefisien mel filterbank, dan 12 jumlah cepstral coefficients. Klasifikasi thresholds diperoleh dua kelas yaitu sehat thresholds < 53,00 dan depresi diatas ≥ 53,00 menggunakan Self-rating Depression Scale. Performa model, akurasi terbaik dicapai RF 0,8481. Pada kelas sehat, presisi dicapai DT 0,8814, recall dan F1-score dicapai RF masing-masing sebesar 0,9706, dan 0,9167. Pada kelas depresi, presisi dicapai RF nilai 0,3333, recall dan F1-score dicapai NB masing-masing sebesar 0,3636, dan 0,2581.

Kata Kunci: Pengenalan pola; Suara; Depresi; Sehat; MFCC

Abstract− The identification of depression patterns from human voices is important because depression can interfere with activities, reduce interest in learning, and hinder socialisation. Depression is a significant problem today because there has been a global increase in the number of people suffering from it. The factors contributing to depression are numerous and complex, and can affect all groups, from children to the elderly. The purpose of this study was to identify depression patterns based on voice feature extraction. The feature extraction method used is Mel-Frequency Cepstral Coefficients (MFCC). The MFCC method is capable of extracting features that closely resemble the human auditory system. The dataset used is the EATD-Corpus, which contains 162 recordings of students from Tongji University in China. The results of the study show that depression and healthy patterns can be distinguished using MFCC parameters, namely 25 measurements per frame, 10 frame intervals, an alpha value of 0.97 as the pre-emphasis coefficient, a maximum of 40 Mel filterbank coefficients, and 12 cepstral coefficients. Classification thresholds can be obtained for two classes: healthy with thresholds < 53.00 and depressed ≥ 53.00 using the Self-Rating Depression Scale.

Keywords: Pattern recognition; Speech; Depressed; Healthy; MFCC

1. PENDAHULUAN

Penyakit depresi pada era komputasi menjadi isu menarik. Penderita depresi saat ini terus mengalami peningkatan, salah satu penyebab yaitu kehadiran perangkat komunikasi [1]. Masalah depresi cukup berdampak secara signifikan terhadap kualitas hidup manusia. Pada negara berkembang seperti Indonesia alat deteksi dini terhadap pasien bergejala depresi [2], tenaga medis, prasarana dan pengobatan masih terbatas.

Permasalahan yang dihadapi pada banyak kasus, bahwa gejala depresi seringkali sulit di diagnosis pada tahap awal [4], [5]. Permasalahan lain, depresi sering kali tidak memiliki gejala fisik yang jelas. Sehingga individu seringkali tidak terdiagnosis tahap awal. Banyak penderita depresi merasa enggan atau malu untuk mengungkapkan [6], [7], [8], hal ini menyebabkan masyarakat tidak dapat melakukan pertolongan dan cenderung membiarkan. Deteksi dini terhadap gejala depresi dan tingkat keakuratan merupakan kunci utama bagi pasien dalam menjalani proses pemulihan [3]. Hal ini membuat pentingnya penerapan teknologi *speech recognition* yang dapat membantu mendeteksi perubahan kecil dalam perilaku individu pada aspek suara.

Masalah depresi cukup berdampak secara signifikan terhadap kualitas hidup manusia. Pada negara berkembang seperti Indonesia alat deteksi dini terhadap pasien bergejala depresi [2], tenaga medis, prasarana dan pengobatan klinis masih terbatas [3], [9]. Tingginya prevalensi penyakit depresi secarSZa global menyebabkan gejala depresi

Submitted: 13 Agustus 2025; Reviewed: 02 September 2025; Accepted; 11 September 2025; Published: 31 Oktober 2025

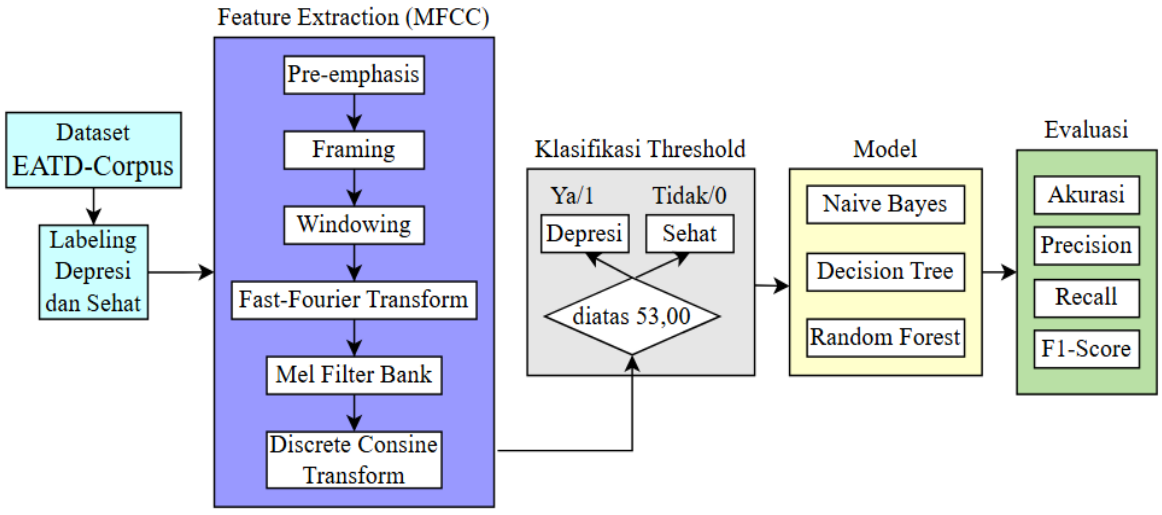
dini terhadap pasien sulit dikenali [4], [5]. Penyakit depresi dapat menyebabkan penderita mengalami emosi [10], fisik tidak stabil, penurunan produktivitas, hubungan sosial [11] sampai potensi bunuh diri [12], [13].
Telah banyak penelitian terkait pasien depresi berdasarkan suara [14], [15], [16], [17], [18], [19], berdasarkan wajah [20], [21], [22], dan berdasarkan pola perilaku [23], [24]. Penelitian Morales [25] menekankan pentingnya penggunaan data audio naturalistik dalam pengembangan sistem deteksi yang lebih akurat dan aplikatif pada dunia nyata. Pengenalan suara depresi dapat dikategorikan dalam fitur prosodi, spectral, dan kualitas suara [26].

Penyelesaikan masalah deteksi gejala depresi berdasarkan suara diperlukan pendekatan secara non-invasif dan inovatif melibatkan *Depression Speech Pattern Recognition* (DSPR). Meskipun kemajuan teknologi bidang DSPR cukup signifikan, sejumlah tantangan masih tetap ada, seperti keterbatasan dataset yang terbuka akibat privasi pasien, keberagaman bahasa, budaya, dan kesulitan dalam validasi klinis. Oleh karena itu, penelitian ini dilakukan penggunaan pendekatan *Mel-Frequency Cepstrum Coefficients* (MFCC). Karena MFCC merupakan metode ekstraksi fitur yang mendekati sistem pendengaran manusia [27], sehingga mampu membedakan pola fitur suara antara depresi dan tidak. Manfaatnya yaitu hasil pola depresi atau tidak dapat digunakan untuk melakukan klasifikasi menggunakan model-model *machine learning* dan *deep learning*.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Blok diagram penelitian ditunjukkan pada Gambar 1. Empat tahap dilakukan, pertama persiapan dataset EATD-Corpus [28]. Satu contoh dari 162 sampel suara ke-11 ditunjukkan pada Gambar 2 dari dataset EATD-Corpus, berisi suara mandarin aksen Tiongkok tradisional telah diterjemahkan. Data EATD-Corpus ada dua file audio WAV bernama *neutral.wav* yaitu audio asli dan *neutral_out.wav* audio telah melalui pembersihan. Kedua melakukan proses pelabelan terhadap 162 suara. Bila *Self-rating Depression Scale* (SDS) melebihi 53,00 maka depresi. Ketiga melakukan ekstraksi fitur menggunakan MFCC. Keempat mengamati pola suara dari ekstraksi fitur hasil dari MFCC [3], [29], [30], dan [31].



Gambar 1. Blok diagram penelitian

Audio ke-11	我的家庭构成啊，额，就是，平常生活和爸爸妈妈住在一起。然后，嗯，爷爷奶奶是分开住，外公去世了，然后，外婆也是分开住，其他还有一些亲戚，不过这些应该不算是家庭构成吧，主要还是就我们家，我一个小孩，然后我的爸爸，我的妈妈，三个人住在一起，家庭构成，家庭构成是什么	Struktur keluarga saya, eh, adalah saya tinggal bersama orang tua saya dalam kehidupan sehari-hari. Lalu, eh, kakek-nenek saya tinggal terpisah, kakek saya meninggal dunia, dan nenek saya juga tinggal terpisah. Ada beberapa kerabat lain, tetapi mereka tidak boleh dianggap sebagai struktur keluarga. Ini terutama keluarga kami, saya seorang anak, dan ayah saya, ibu saya, kami bertiga tinggal bersama. Bagaimana struktur keluarga itu?

Gambar 2. Terjemahan suara sampel ke-11

https://doi.org/10.33998/processor.2025.20.2.2513
Submitted: 13 Agustus 2025; Reviewed: 02 September 2025; Accepted; 11 September 2025; Published: 31 Oktober 2025

2.2 Dataset

EATD-Corpus berisi rekaman audio format WAV dan transkrip teks dari wawancara 162 peserta bersatus mahasiswa Universitas Tongji Tiongkok. Durasi total rekaman audio pada EATD-Corpus sekitar 2,26 jam. Mahasiswa yang direkam telah menyetujui dan menjamin keaslian audio. Masing-masing mahasiswa diwajibkan menjawab tiga pertanyaan yang dipilih secara acak. Setelah mengisi rekaman para mahasiswa mengisi kuesioner SDS. Kuesioner SDS berisi 20 pertanyaan menilai empat parameter karakteristik gejala penyakit depresi: efek pervasif, efek fisiologis, gangguan lain, dan aktivitas psikomotorik [32]. SDS merupakan kuesioner yang umum digunakan psikolog dalam memilah individu berpotensi depresi pada praktik klinis. Banyak penelitian telah membahas parameter yang efektif dalam membedakan gejala depresi atau tidak dengan fitur spektral dan MFCC [29], [30], [31], [33], [34]. MFCC mampu membedakan suara termasuk fitur prosodi (nama, energi suara, *formant*) atau fitur spektral (MFCC, LPCC, GFCC). Fitur prosodi berhubungan dengan saluran vokal dan model suara seseorang, seperti variasi, ukuran saluran vokal, gerakan artikulator, dan kecepatan bicara. Fitur spektral berfokus pada karakterisasi sinyal, frekuensi, yang dapat menambah informasi berguna untuk fitur prosidi [35].

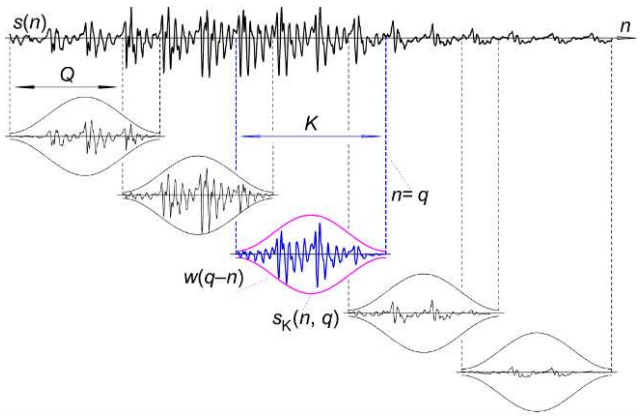
2.3 Mel-Frequency Cepstral Coefficients

MFCC dapat digunakan untuk ekstraksi fitur suara depresi, perhitungan MFCC ada enam tahap.*Pre-emphasis* melakukan proses peningkatan frekuensi tinggi dari sinyal suara sumber. Fungsi pertama, memampatkan efek rekaman suara yang cenderung melemahkan frekuensi tinggi. Kedua meningkatkan rasio terhadap *noise* pada frekuensi tinggi. Ketiga, meningkatkan akurasi suara, karena fitur suara frekuensi tinggi dapat lebih jelas. Keempat, menyeimbangkan spektrum sebelum ekstraksi fitur, dalam hal ini MFCC. Persamaan 1 merupakan perhitungan *pre-emphasis* [27], [29], dinotasikan p(n) sebagai sinyal output, s(n) sebagai sinyal input asli, dan α sebagai koefisien antara 0,9 hingga 1,0.

$$p(n) = s(n) - \alpha s(n-1);\ 0{,}9 \leq \alpha \leq 1{,}0 \tag{1}$$

Framing memecah sinyal menjadi potongan kecil yang disebut *frame* dengan asumsi sinyal bersifat *stationer*. Pada tahap *framing* sinyal suara dibagi menjadi beberapa *frame*, terdiri dari N sampel. *Frame* yang berdekatan dipisahkan sejauh M (M\N). Analisis sinyal suara dalam *frame* biasanya 20 – 40 ms [28], namun penelitian [29] lebih memilih 10 – 30 ms. *Frame* diambil sepanjang mungkin guna mendapatkan resolusi frekuensi terbaik, waktu dipilih sependek mungkin bertujuan mendapatkan ranah waktu terbaik. *Overlapping* pada sebuah sinyal ditunjukkan pada Gambar 3.
Hamming window bertujuan menjaga kesinambungan ujung pertama dan terakhir setiap *frame* menjadi nol. Setiap *frame* harus dikalikan dengan *hamming window*, sehingga dapat meminimalkan gangguan ketika proses *framing*. Sinyal suara merupakan sinyal nyata, sehingga memiliki waktu yang terbatas. Perhitungan *hamming windowing* ditunjukkan persamaan 2 [27], [29], w(n)merupakan nilai fungsi frame pada indeks ke-n, dan N adalah panjang frame (jumlah sampel setiap *frame*).

$$w(n) = 0{,}54 - 0{,}46 \cdot cos\left(\frac{2\pi n}{N-1}\right)\ untuk\ 0 \leq n \leq N-1 \tag{2}$$



Gambar 3. *Overlapping* antar *frame* pada sebuah sinyal

Fast-Fourier Transform (FFT) yaitu menstransformasikan masing-masing *frame* dari N sampel satuan waktu ke satuan frekuensi [29]. FFT pada data audio digunakan untuk menganalisis spektrum frekuensi pada masing-

https://doi.org/10.33998/processor.2025.20.2.2513
Submitted: 13 Agustus 2025; Reviewed: 02 September 2025; Accepted; 11 September 2025; Published: 31 Oktober 2025

masing *frame* sinyal suara, mengetahui banyaknya energi yang terkandung pada setiap frekuensi, dan menyediakan data ketika proses *Mel Filterbank*. Hasil FFT merupakan deretan bilangan kompleks, masing-masing menunjukkan amplitudo dan fase frekuensi tertentu. Perhitungan FFT ditunjukkan pada persamaan 3 [27], notasi f(n) menyatakan sinyal input dalam satuan waktu setiap *frame*, W_k menyatakan spektrum frekuensi, N menyatakan panjang jumlah sampel *frame*, j menyatakan unit imajiner, dan $e^{-2\pi j\frac{kn}{n}}$ menyatakan eksponensial kompleks yang mendefinisikan transformai terbalik.

$$f(n) = \sum_{K=0}^{N-1} W_k e^{\frac{-2\pi j k n}{N}}; 0 \leq n \leq N-1 \quad (3)$$

Mel Filterbank merupakan kumpulan filter berbentuk segitiga, disusun berdasarkan skala *Mel* supaya mendekati suara telinga manusia, sehingga lebih sensitif terhadap frekuensi rendah dari pada tinggi [3], [29]. Resolusi frekuensi telinga manusia tidak mengikuti skala linear pada seluruh spektrum audio. Sehingga setiap frekuensi yang diukur dalam satuan Hz, nada subyektif diukur pada skala *Mel* [29]. Skala frekuensi *Mel* memiliki jarak linear kurang dari 1000 Hz dan logaritmik diatas 1000 Hz dan filter memiliki bentuk segitiga [3], [29]. Menghitung *Mel* digunakan persamaan 4, notasi f menyatakan frekuensi asli satuan Hz, $\frac{f}{700}$ menyatakan proses normalisasi frekuensi, $1+\frac{f}{700}$ menyatakan mencegah log(0) dan memberikan skala yang stabil, log_{10} mengubah ke skala logaritmik sesuai persepsi pendengaran, dan 2595 menyatakan konstanta.

$$Mel(f) = 2595 \cdot log_{10}\left(1+\frac{f}{700}\right) \quad (4)$$

DCT digunakan untuk mengkompresi suara log-energi *filterbank* menjadi sejumlah koefisien lalu menghilangkan korelasi antar fitur supaya suara lebih mudah digunakan, dan memfokuskan data suara penting pada koefisien awal, misalnya digunakan 12-13 dari sejumlah 26-40. Telinga manusia merespon energi suara secara logaritmik [35], sehingga perlu meniru persepsi manusia terhadap intensitas suara. Energi sinyal suara dapat tinggi atau rendah, fungsi log untuk menyusutkan skala energi supaya tidak terlalu ekstrem dan membuat pola suara lebih stabil. Hasil dari *logarithm* digunakan sebagai input proses DCT, karena DCT bekerja lebih baik menggunakan nilai yang merata [29]. Perhitungan DCT ditunjukkan pada persamaan 5 [27], notasi C_j menyatakan koefisien DCT ke-j (hasil tranformasi), X_i menyatakan jumlah masukan ke-i dalam domain waktu pada data asli, M menyatakan jumlah total sampel masukan (panjang array sinyal), j menyatakan indeks komponen frekuensi yang dihitung biasanya $0 \leq j \leq M-1$, dan $\cos(j(i-1)/2\frac{\pi}{M})$ menyatakan fungsi dasar DCT, fungsi ini menentukan seberapa besar cosinus pada frekeunsi j pada sinyal masukan.

$$C_j = \sum_{i=1}^{M} X_i \cdot \cos\left(j\,(i-1)/2\frac{\pi}{M}\right) \quad (5)$$

2.4 Model Yang Digunakan

Beberapa model yang sering digunakan yaitu SVM [36], [37], [38], *Random Forest* [36], [38], *Decision Tree*, kNN [37], [38], [39], *Naïve Bayes* [38], dan *Logistic Regression* [40]. Masing-masing model memiliki keunggulan, namun pada penelitian ini digunakan tiga model yaitu *Naïve Bayes* (NB), *Decision Tree* (DT), dan *Random Forest* (RF). *Naïve Bayes* termasuk algoritma probabilistik. Dinamakan *Naïve* karena diasumsikan sederhana namun cepat. *Naïve Bayes* digunakan ketika tujuannya untuk memberikan label pada data input berdasarkan fitur-fiturnya [38]. Konsep perhitungan *Naïve Bayes* tetap sama, namun implementasi bergantung pada jenis data yaitu kontinyu atau diskret, jenis fitur (suara, teks, citra), dan tujuan (klasifikasi, prediksi, probabilitas). Pada paper ini digunakan *Gaussian Naïve Bayes* karena fitur suara seperti MFCC bersifat kontinyu.

Decision Tree merupakan model terbimbing, banyak digunakan dalam klasifikasi. *Decision Tree* adalah struktur pohon dimulai dari simpul akar, kemudian bercabang menjadi simpul. Setiap simpul mewakili kelas mengarah ke hasil. Ide dasar dari *Decision Tree* yaitu memisahkan seluruh data sampel menjadi beberapa sub-sampel berdasarkan kriteria tertentu. Variabel target dapat menentukan jenis pohon. Variabel input dan output berupa kategorikal atau kontinyu. Hal ini diklasifikasikan menjadi dua jenis. Ketika variabel target bersifat kategorikal, disebut *Decision Tree* kategorikal. Ketika variabel target bersifat kontinyu, disebut *Decision Tree* kontinyu [38].

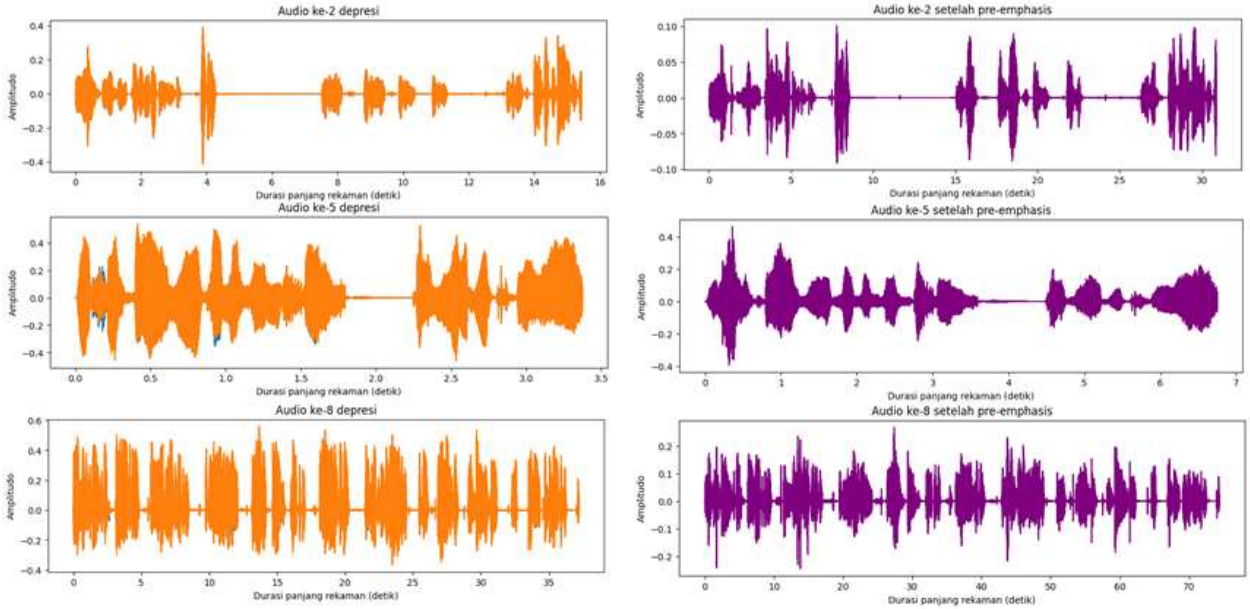
Random Forest merupakan model pembelajaran *ensemble*, terdiri dari sejumlah besar pohon keputusan untuk membuat prediksi lebih akurat. Setiap pohon keputusan individu dapat menghasilkan prediksi kelas, selanjutnya kelas suara terbaik menjadi prediksi akhir. *Random Forest* bekerja melibatkan banyak pohon

https://doi.org/10.33998/processor.2025.20.2.2513
Submitted: 13 Agustus 2025; Reviewed: 02 September 2025; Accepted; 11 September 2025; Published: 31 Oktober 2025

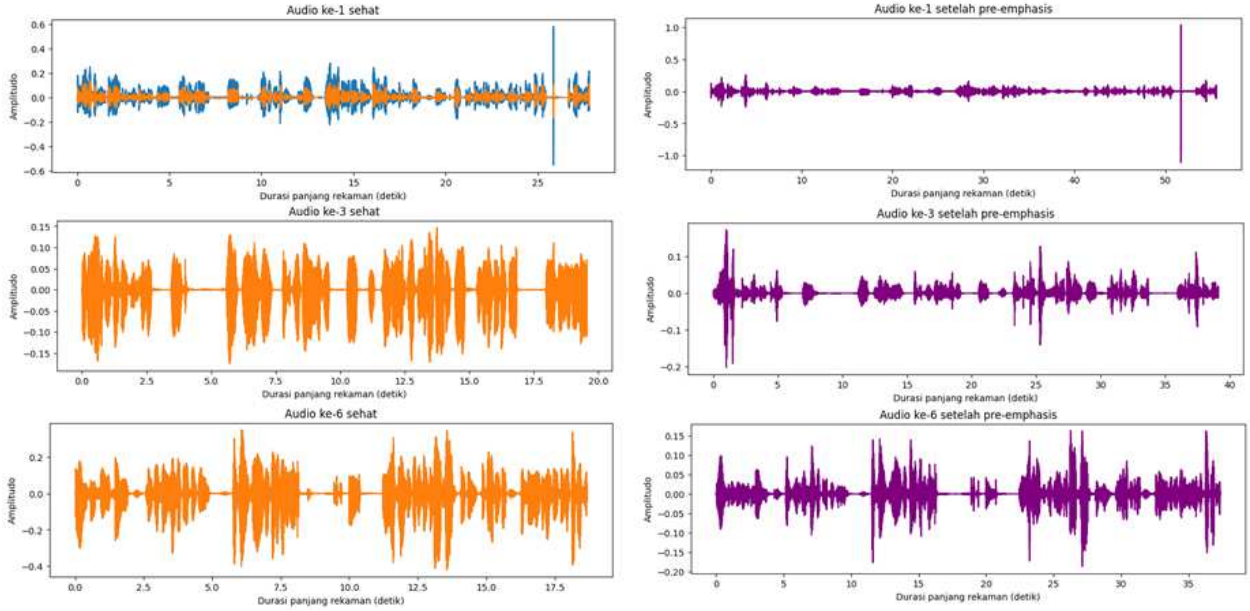
keputusan. Setiap pohon keputusan dibentuk dari sampling data dan semua fitur. Semua pohon yang terbentuk akan dilakukan *voting* guna menentukan kelas [41].

3. HASIL DAN PEMBAHASAN

Penelitian ini mengambil contoh enam sampel suara dari dataset EATD-Corpus, masing-masing tiga sampel depresi dan tiga sampel sehat ditunjukkan pada Gambar 4 dan 5. Setelah data terlabel, kemudian dicari polanya menggunakan MFCC. Parameter-parameter yang digunakan yaitu *frame_size* = 25 sebagai ukuran masing-masing *frame*, *frame_stride* = 10 sebagai jarak antar *frame*, *Alpa* = 0,97 sebagai nilai koefisien *pre-emphasis*, M = 40 sebagai jumlah maksimum koefisien *mel filterbank*, *num_ceps* = 12 sebagai jumlah *cepstral coefficients* yang digunakan untuk model klasifikasi. Parameter-parameter tersebut digunakan untuk menghasilkan ekstraksi fitur dari MFCC, kemudian fitur tersebut dicari nilai rata-rata untuk dikenali ciri depresi atau tidak [27].
Gambar 4 menunjukkan sampel audio stereo depresi sebelum dan sesudah *pre-emphasis*, warna biru dan oranye kemungkinan salah satu suara kiri atau kanan. Gambar 5 menunjukkan sampel audio stereo sehat sebelum dan setelah *pre-emphasis*. Sumbu X mewakili durasi waktu milidetik. Sumbu Y menunjukkan amplitudo.



Gambar 4. Bentuk gelombang audio depresi sebelum dan setelah proses *pre-emphasis*.



Gambar 5. Bentuk gelombang audio sehat sebelum dan setelah proses *pre-emphasis*.

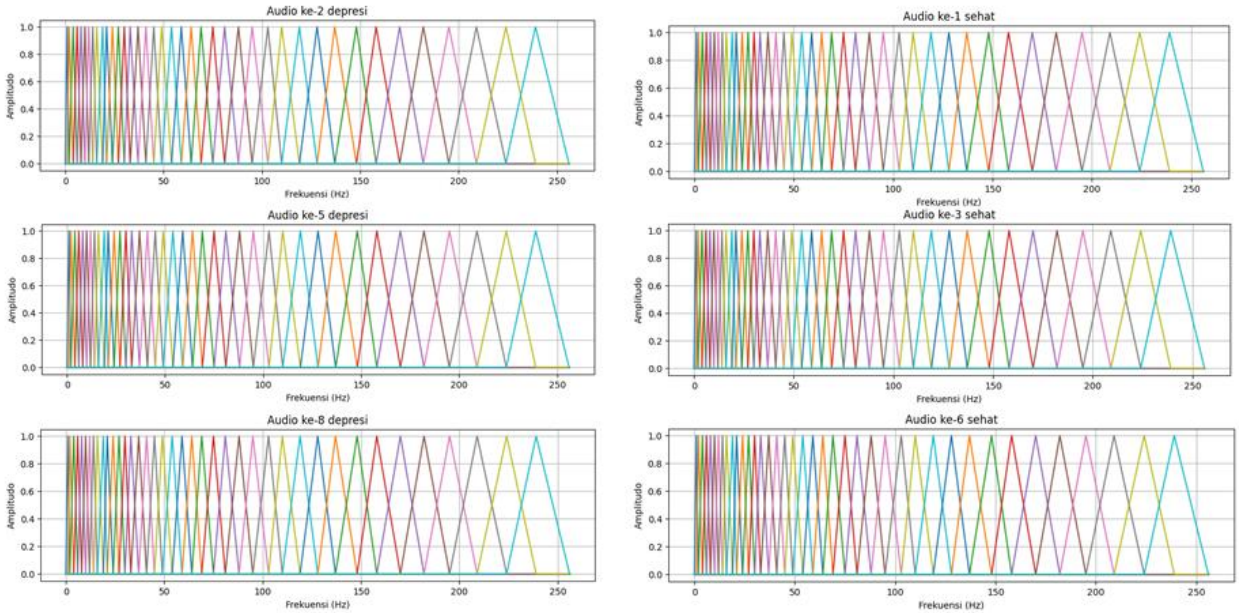
https://doi.org/10.33998/processor.2025.20.2.2513
Submitted: 13 Agustus 2025; Reviewed: 02 September 2025; Accepted; 11 September 2025; Published: 31 Oktober 2025

Hasil *pre-emphasis* Gambar 4 dan 5 menunjukkan *filtering* sehingga frekuensi tinggi muncul, dan frekuensi rendah telah berkurang atau hilang. Kemudian frekuensi tinggi tidak dominan. Tahap *pre-emphasis* menggali banyak informasi penting dari pada data audio asli, sehingga hasil dari *pre-emphasis* dapat digunakan sebagai masukkan tahap *framing*.

Tujuan *framing* yaitu pertama, membagi sinyal suara menjadi beberapa potongan yang disebut *frame* dengan durasi tetap. Kedua supaya sinyal dapat bersifat *stationer* atau stabil dalam setiap *frame*, karena suara manusia bersifat *non-stationer* atau tidak stabil. Ketiga, memberikan overlap untuk mencegah kehilangan informasi antar frame. Suara manusia dapat dianggap stabil pada durasi tertentu, sehingga pada penelitian ini menggunakan ukuran 25ms (0,025 detik), karena sinyal dianggap *stationer*. Kemudian jarak antar *frame* ditentukan 10ms (0,01 detik), bertujuan membantu kontinuitas sinyal tetap terjaga. Dengan demikian *frame* saling *overlap* sebesar 15ms, berasal dari 25ms – 10ms = 15ms.

Hamming window adalah fungsi yang digunakan dalam pemrosesan sinyal untuk memusatkan energi *frame* pada spektrum, dan mengurangi disparitas tepi sinyal. Diasumsikan ketika ada gelombang sinus yang terpotong tiba-tiba maka sinyal audio akan menghasilkan suara hentakan mendadak. Oleh sebab itu, teknik *hamming window* digunakan untuk memperhalus tepi *frame*. Sehingga membuat sinyal pada tepi awal dan akhir menjadi nol secara perlahan, audio tidak berhenti tiba-tiba. Dengan demikian hasil analisis frekuensi menjadi lebih bersih dan akurat, hal ini sangat penting sebelum FFT supaya hasil fitur suara mewakili suara asli.

FFT bertujuan mengubah *frame-frame* kecil berdomain waktu menjadi representasi domain frekuensi (*spectrum*). FFT adalah salah satu algoritma yang cepat untuk menerapkan *Discrete Fourier Transform* (DCT) yang beroperasi pada sinyal diskrit [27]. Hal ini membantu melihat frekuensi mana yang ada dan seberapa kuat frekuensi dalam *frame* tersebut. Selama proses, FFT dilakukan terhadap semua *frame* dari sinyal yang sudah melalui tahap *windowing*.



Gambar 6. *Mel filterbank*

Mel Filterbank digunakan untuk meniru cara telinga manusia dalam menangkap suara. Oleh sebab itu diperlukan *mel filterbank*, untuk menyaring sinyal dari FFT supaya fokus pada pola energi suara dalam rentang frekuensi tertentu. Hasil dari *mel filterbank* akan digunakan sebagai masukan tahap DCT untuk memperoleh ekstraksi fitur dengan MFCC. Gambar 6 terdapat 40 segitiga karena menggunakan *nfilt*=40. Filter di frekuensi rendah rapat banyak *overlap* pada sisi kiri. Filter di frekuensi tinggi tidak *overlap* atau lebih lebar. Sumbu X bin indek menunjukkan indek dari hasil FFT. Setiap angka pada sumbu X mewakili frekuensi diskrit dalam bentuk bin (*bucket*), misal bin ke-0, bin ke-50, bin ke-100 dan seterusnya bukan dalam satuan Hz. Sumbu Y adalah amplitude mewaliki nilai bobot dari filter segitiga setiap bin. Nilai rentang 0 sampai 1, pada puncak segitiga filter memiliki nilai maksimum dan menurun sampai 0 dibawah, mulai dari sisi kiri ke arah kanan.

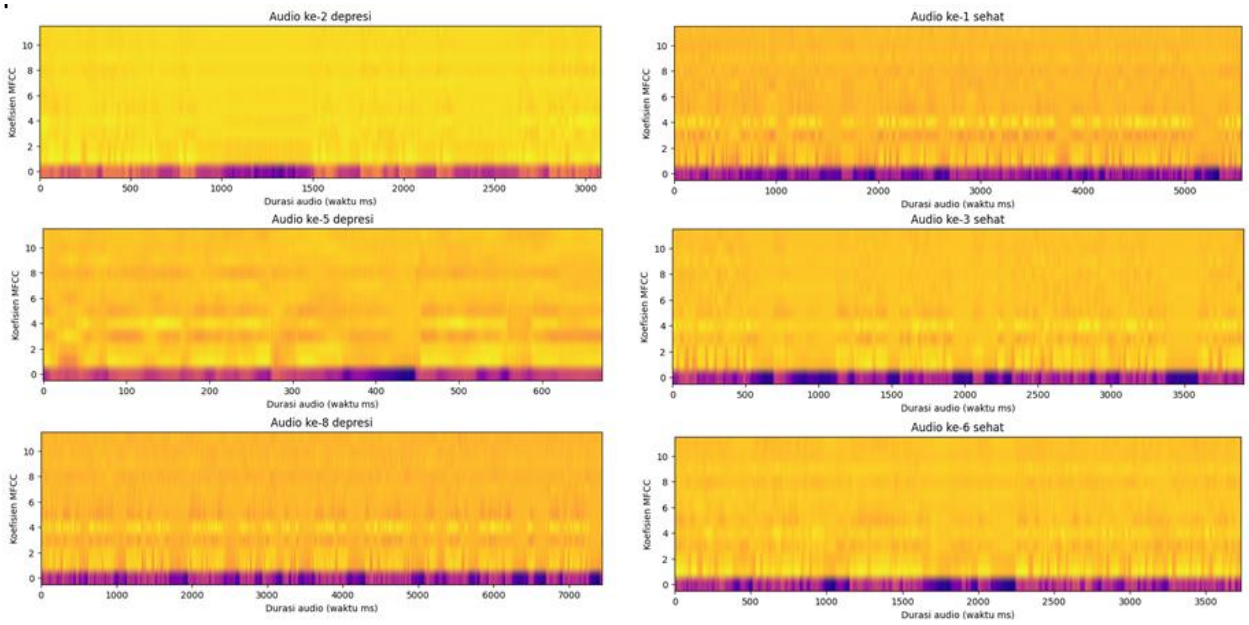
MFCC bertujuan menangkap karakteristik suara manusia dengan cara seperti telinga manusia ketika mendengarkan suara. MFCC mengurangi informasi yang tidak penting dalam sinyal untuk keperluan klasifikasi, prediksi, pengenalan suara atau analisis suara. MFCC dapat menangkap fitur timbre dan fonetik suara dan mengkompresi isi frekuensi setiap *frame* menggunakan DCT untuk mendapatkan vektor yang ringkas. Proses

https://doi.org/10.33998/processor.2025.20.2.2513
Submitted: 13 Agustus 2025; Reviewed: 02 September 2025; Accepted; 11 September 2025; Published: 31 Oktober 2025

memperoleh ekstraksi fitur ini menggunakan program Python 3.11 dengan teks editor *Visual Studio Code*. Perangkat yang digunakan *Intel Core-2-duo* RAM 4 GB dengan penyimpan SSD-250GB.

Hasil ekstraksi fitur bentuk spektogram dari sampel depresi dan sehat ditunjukkan pada Gambar 7 ada representasi warna menunjukkan nilai atau kekuatan setiap koefisien MFCC dalam suatu *frame*. Warna kuning menyatakan nilai lebih tinggi dan warna ungu menyatakan nilai lebih rendah. Sumbu X menjelaskan indek *frame* dalam milidetik, yaitu suara dibagi menjadi *frame-frame* kecil. Sumbu Y sebagai koefisien *cepstral*, merupakan indek koefisien MFCC biasanya dalam 12 atau 13 pertama yang ditampilkan. Koefisien ini menunjukkan spektrum suara berubah di setiap *frame*. MFCC melakukan ekstraksi fitur menjadi 12 koefisien dimulai dari MFCC-0 hingga MFCC-11 [42].

.



Gambar 7. Spektogram MFCC

Selanjutnya dari spektogram MFCC Gambar 7 dapat direpresentasikan menjadi MFCC numerik seperti Tabel 1. Setiap baris mewakili sampel rekaman audio sebanyak 162, kolom 0 – 11 berisi MFCC rata-rata untuk masing-masing koefisien. Kolom KT menunjukkan klasifikasi *threshold*, angka ini telah ditentukan dan dianotasi oleh psikolog atau psikiater berdasarkan kuisioner *Self-rating Depression Scale* [32]. Kuisioner berisi 20 pertanyaan dengan empat parameter karakteristik gejala depresi [28]. Kolom terakhir adalah label kelas yang menunjukkan kondisi audio berisi depresi bila rekaman berasal dari subyek depresi atau sehat bila berasal dari subyek sehat.

Tabel 1. Representasi MFCC numerik

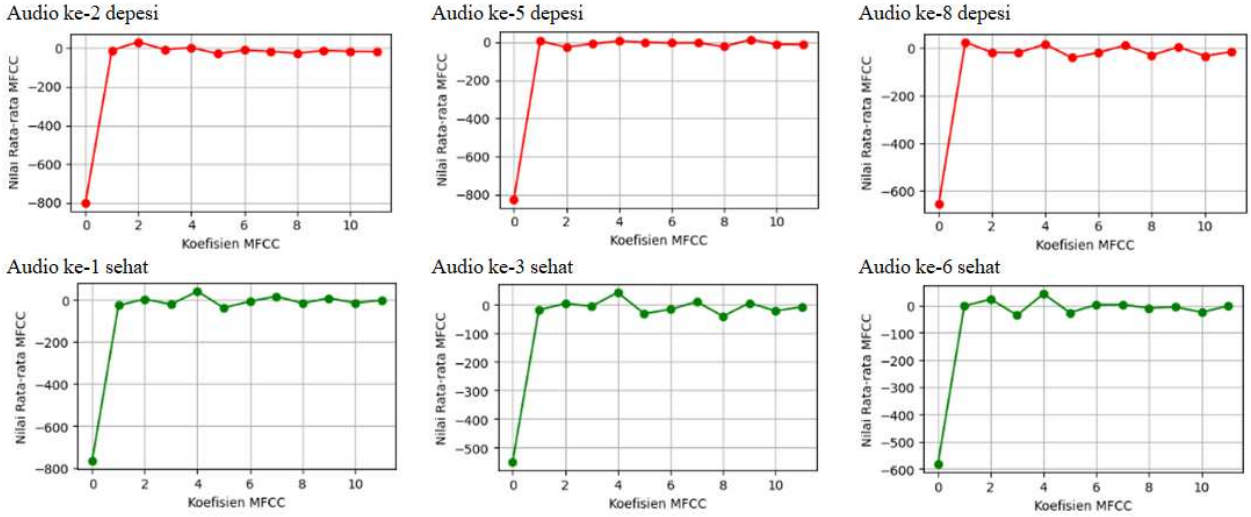
Data ke-	MFCC[0]	MFCC[1]	MFCC[2]	…	MFCC[10]	MFCC[11]	Hasil Thresholds	Label
1	-764,355	-26,797	3,050	…	-15,387	-2,347	52,50	Sehat
2	-799,562	-10,609	31,938	…	-16,524	-17,652	82,50	Depresi
3	-799,562	-10,609	31,938	…	-22,544	-8,207	51,25	Sehat
4	-799,562	-10,609	31,938	…	-10,729	-12,044	56,25	Depresi
5	-652,740	26,618	-17,258	…	-31,719	-14,255	68,75	Depresi
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
160	-796,354	-71,085	-56,076	…	-2,965	-16,939	65,00	Depresi
161	-644,905	-15,200	-3,842	…	-6,055	-14,815	32,50	Sehat
162	-699,680	37,617	-28,778	…	-6,781	-16,023	41,25	Sehat

Gambar 8 berisi enam visualisasi masing-masing tiga depresi warna merah dan tiga sehat warna hijau. Tanda titik menginterpretasikan nilai rata-rata 12 koefisien MFCC untuk masing-masing kelas sehat dan depresi. Garis warna merah menunjukkan tiga sampel ke-2, ke-5, ke-8 dari suara depresi, dan garis warna hijau menunjukkan tiga sampel ke-1, ke-3, ke-6 dari suara sehat. Setiap subplot memperlihatkan nilai rata-rata 12 koefisien MFCC, yang merepresentasikan spektrum suara dalam domain *log-mel* dan digunakan untuk menganalisis karakteristik suara manusia.

https://doi.org/10.33998/processor.2025.20.2.2513
Submitted: 13 Agustus 2025; Reviewed: 02 September 2025; Accepted; 11 September 2025; Published: 31 Oktober 2025

Garis merah menunjukkan pola amplitudo yang lebih rendah dan relatif rata atau datar pada beberapa titik koefisien dibandingkan kategori sehat. Terlihat pada audio ke-1 dan ke-2, beberapa nilai MFCC seperti koefisien ke-2 hingga ke-8 lebih stabil atau berfluktuasi kecil, hal ini mencerminkan pola suara yang monoton atau kurang ekspresif. Koefisien awal MFCC[0] hingga MFCC[3] cenderung rendah dibandingkan dengan yang sehat, ada kemungkinan berkaitan dengan perubahan frekuensi dasar.

Garis hijau memiliki pola MFCC lebih bervariatif dan fluktiatif pada ketiga audio. Nilai koefisien pada MFCC[1], MFCC[2], MFCC[3], dan MFCC[7] lebih tinggi. Hal ini menunjukkan kemungkinan ada dinamika vokal lebih kuat atau artikulasi lebih jelas. Pola yang naik turun secara tajam menandakan ada perubahan frekuensi dan intensitas suara lebih dinamis, ini merupakan ciri umum suara normal atau orang sehat.



Gambar 8. Ekstraksi fitur MFCC sehat dan depresi

Setelah memperoleh ekstraksi fitur, selanjutnya setiap *file* suara direpresentasikan dalam bentuk numerik berupa koefisien yang mencerminkan karakteristik akustik dari suara tersebut. Tabel 1 digunakan sebagai input untuk proses ketiga model NB, DT, dan RF dalam mencari performa akurasi, *precision*, *recall*, dan F1-*score*. Tabel 2 dan 3 menunjukkan evaluasi kelas sehat dan depresi. Tampak pada kelas sehat, *recall* dan F1-*score* tertinggi model RF, sedangkan *precision* pada model DT. Untuk kelas depresi *recall* dan F1-*score* tertinggi pada model NB, sedangkan *precision* pada model RF.

Tabel 2. Evaluasi pada kelas sehat

Model	***Precision***	***Recall***	***F1-score***
NB	0,8696	0,8824	0,8759
DT	**0,8814**	0,7647	0,8189
RF	0,8684	**0,9706**	**0,9167**

Tabel 3. Evaluasi pada kelas depresi

Model	***Precision***	***Recall***	***F1-score***
NB	0,2000	**0,3636**	**0,2581**
DT	0,2000	0,1818	0,1905
RF	**0,3333**	0,0909	0,1429

Tabel 4 memperlihatkan akurasi ketiga model pada dua kelas sehat dan depresi. Akurasi tertinggi dicapai model RF. Kemudian pada *support* tampak terlihat distribusi kelas tidak seimbang, dengan 68 individu sehat dan 11 depresi. Hal ini dapat mempengaruhi model dalam menangani kelas minoritas yaitu depresi.

Tabel 4. Akurasi ketiga model

Model	**Akurasi**	***Support***	
		Sehat	Depresi
DT	0,7089	68	11
NB	0,7848	68	11
RF	**0,8481**	68	11

https://doi.org/10.33998/processor.2025.20.2.2513
Submitted: 13 Agustus 2025; Reviewed: 02 September 2025; Accepted; 11 September 2025; Published: 31 Oktober 2025

4. KESIMPULAN

Hasil penelitian menunjukkan pola depresi berhasil terlihat menggunakan parameter MFCC yaitu 25 ukuran masing-masing *frame*, 10 jarak antar *frame*, alpha 0,97 sebagai nilai koefisien *pre-emphasis*, 40 jumlah maksimum koefisien *mel filterbank*, dan 12 jumlah *cepstral coefficients*. Klasifikasi *thresholds* dapat diperoleh dua kelas yaitu sehat dengan *thresholds* < 53,00 dan depresi diatas ≥ 53,00 menggunakan *Self-rating Depression Scale*. Penelitian ini menggunakan metode MFCC untuk menemukan ekstraksi fitur suara depresi dan sehat. Secara visual, grafik yang dihasilkan mengindikasikan ada perbedaan karakteristik suara antara orang depresi dan sehat. Suara orang depresi cenderung memiliki pola yang lebih datar dan stabil, sedangkan suara sehat lebih berfluktuasi dan kaya secara spektral. Hal ini mendukung penggunaan MFCC sebagai fitur utama dalam mendeteksi kondisi psikologis orang berdasarkan suara.

Performa model, akurasi terbaik dicapai RF sebesar 0,8481. Pada kelas sehat, presisi dicapai model DT dengan nilai 0,8814, untuk *recall* dan *F1-score* dicapai model RF masing-masing sebesar 0,9706, dan 0,9167. Pada kelas depresi presisi dicapai model RF dengan nilai 0,3333, untuk *recall* dan *F1-score* dicapai model NB masing-masing sebesar 0,3636, dan 0,2581.

Dataset yang digunakan dalam model ini memiliki keterbatasan yaitu ukuran total dataset relatif kecil, hal ini dapat membatasi kemampuan model dalam mengenal pola keseluruhan. Selain itu, terdapat potensi bias bahasa karena dataset berasal dari satu sumber suara bahasa Tiongkok. Hal ini dapat mengakibatkan penurunan kinerja model ketika diterapkan. Oleh sebab itu penelitian mendatang perlu mempertimbangkan keberagaman suara untuk meningkatkan generalisasi model.

Salah satu kendala penelitian ini yaitu ketidakseimbangan data antar kelas yang dapat mempengaruhi kinerja model. Rencana penelitian lanjutan perlu melibatkan teknik *Synthetic Minority Oversampling Technique* (SMOTE) untuk mengatasi data tidak seim3Ebang dengan menghasilkan sampel sintetis dari kelas minoritas. Diharapkan menggunakan pendekatan ini, distribusi data lebih seimbang dan model dapat lebih adil terhadap semua kelas.

REFERENCES

[1] Y. Zhang *et al.*, "Identifying Depression-Related Topics in Smartphone-Collected Free-Response Speech Recordings Using an Automatic Speech Recognition System and A Deep Learning Topic Model," *J. Affect. Disord.*, vol. 355, no. September 2023, hal. 40–49, 2024, doi: 10.1016/j.jad.2024.03.106.

[2] L. Munira, P. Liamputtong, dan P. Viwattanakulvanid, "Barriers and facilitators to access mental health services among people with mental disorders in Indonesia: A qualitative study," *Belitung Nurs. J.*, vol. 9, no. 2, hal. 110–117, 2023, doi: 10.33546/bnj.2521.

[3] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, dan A. Othmani, "MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech," *Biomed. Signal Process. Control*, vol. 71, no. PA, hal. 103107, 2022, doi: 10.1016/j.bspc.2021.103107.

[4] R. Cheung, S. O'Donnell, N. Madi, dan E. M. Goldner, "Factors associated with delayed diagnosis of mood and/or anxiety disorders," *Heal. Promot. Chronic Dis. Prev. Canada*, vol. 37, no. 5, hal. 137–148, 2017, doi: 10.24095/hpcdp.37.5.02.

[5] R. Huerta-Ramírez, J. Bertsch, M. Cabello, M. Roca, J. M. Haro, dan J. L. Ayuso-Mateos, "Diagnosis delay in first episodes of major depression: A study of primary care patients in Spain," *J. Affect. Disord.*, vol. 150, no. 3, hal. 1247–1250, 2013, doi: 10.1016/j.jad.2013.06.009.

[6] L. J. Barney, K. M. Griffiths, A. F. Jorm, dan H. Christensen, "Stigma about depression and its impact on help-seeking intentions," *Aust. N. Z. J. Psychiatry*, vol. 40, no. 1, hal. 51–54, 2006, doi: 10.1111/j.1440-1614.2006.01741.x.

[7] E. Samari *et al.*, "Perceived mental illness stigma among family and friends of young people with depression and its role in help-seeking: a qualitative inquiry," *BMC Psychiatry*, vol. 22, no. 1, hal. 1–13, 2022, doi: 10.1186/s12888-022-03754-0.

[8] R. M. Epstein *et al.*, "'I didn't know what was wrong:' How people with undiagnosed depression recognize, name and explain their distress," *J. Gen. Intern. Med.*, vol. 25, no. 9, hal. 954–961, 2010, doi: 10.1007/s11606-010-1367-0.

[9] F. Farajullah dan M. Murinto, "Sistem Pakar Deteksi Dini Gangguan Kecemasan (Anxiety) Menggunakan Metode Forward Chaining Berbasis Web," *JSTIE (Jurnal Sarj. Tek. Inform.*, vol. 7, no. 1, hal. 1, 2019, doi: 10.12928/jstie.v7i1.15800.

[10] M. Kapitány-Fövény, M. Vetró, G. Révy, D. Fabó, D. Szirmai, dan G. Hullám, "EEG based depression detection by machine learning: Does inner or overt speech condition provide better biomarkers when using emotion words as experimental cues?," *J. Psychiatr. Res.*, vol. 178, no. January, hal. 66–76, 2024, doi: 10.1016/j.jpsychires.2024.08.002.

[11] H. Kim, S. H. Lee, S. E. Lee, S. Hong, H. J. Kang, dan N. Kim, "Depression prediction by using ecological momentary assessment, actiwatch data, and machine learning: Observational study on older adults living alone," *JMIR mHealth uHealth*, vol. 7, no. 10, 2019, doi: 10.2196/14149.

[12] T. M. H. Li *et al.*, "Detection of Suicidal Ideation in Clinical Interviews for Depression Using Natural Language

https://doi.org/10.33998/processor.2025.20.2.2513
Submitted: 13 Agustus 2025; Reviewed: 02 September 2025; Accepted; 11 September 2025; Published: 31 Oktober 2025

Processing and Machine Learning: Cross-Sectional Study," *JMIR Med. Informatics*, vol. 11, no. 1, hal. 1–13, 2023, doi: 10.2196/50221.
[13] Bahman Zohuri dan Siamak Zadeh, "The Utility of Artificial Intelligence for Mood Analysis, Depression Detection, and Suicide Risk Management," *J. Heal. Sci.*, vol. 8, no. 2, hal. 67–73, 2020, doi: 10.17265/2328-7136/2020.02.003.
[14] P. Lopez-Otero, L. Docio-Fernandez, A. Abad, dan C. Garcia-Mateo, "Depression detection using automatic transcriptions of de-identified speech," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2017-Augus, hal. 3157–3161, 2017, doi: 10.21437/Interspeech.2017-1201.
[15] Z. Liu, D. Wang, L. Zhang, dan B. Hu, "A Novel Decision Tree for Depression Recognition in Speech," 2020, [Daring]. Tersedia pada: http://arxiv.org/abs/2002.12759
[16] B. Sumali *et al.*, "Speech quality feature analysis for classification of depression and dementia patients," *Sensors (Switzerland)*, vol. 20, no. 12, hal. 1–17, 2020, doi: 10.3390/s20123599.
[17] E. Villatoro-Tello, S. P. Dubagunta, J. Fritsch, G. Ramírez-De-La-Rosa, P. Motlicek, dan M. Magimai.-Doss, "Late fusion of the available lexicon and raw waveform-based acoustic modeling for depression and dementia recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 1, hal. 161–165, 2021, doi: 10.21437/Interspeech.2021-1288.
[18] O. Simantiraki, P. Charonyktakis, A. Pampouchidou, M. Tsiknakis, dan M. Cooke, "Glottal source features for automatic speech-based depression assessment," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2017-Augus, hal. 2700–2704, 2017, doi: 10.21437/Interspeech.2017-1251.
[19] N. Seneviratne dan C. Espy-Wilson, "Speech based depression severity level classification using a multi-stage dilated CNN-LSTM model," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2, hal. 746–750, 2021, doi: 10.21437/Interspeech.2021-1967.
[20] S. A. Nasser, I. A. Hashim, dan W. H. Ali, "A review on depression detection and diagnoses based on visual facial cues," *2020 3rd Int. Conf. Eng. Technol. its Appl. IICETA 2020*, hal. 35–40, 2020, doi: 10.1109/IICETA50496.2020.9318860.
[21] L. R. Demenescu, R. Kortekaas, J. A. den Boer, dan A. Aleman, "Impaired attribution of emotion to facial expressions in anxiety and major depression," *PLoS One*, vol. 5, no. 12, 2010, doi: 10.1371/journal.pone.0015058.
[22] X. Zhou, K. Jin, Y. Shang, dan G. Guo, "Visually Interpretable Representation Learning for Depression Recognition from Facial Images," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, hal. 542–552, 2020, doi: 10.1109/TAFFC.2018.2828819.
[23] M. Teychenne, K. Ball, dan J. Salmon, "Sedentary behavior and depression among adults: A review," *Int. J. Behav. Med.*, vol. 17, no. 4, hal. 246–254, 2010, doi: 10.1007/s12529-010-9075-z.
[24] M. C. Lovejoy, P. A. Graczyk, E. O'Hare, dan G. Neuman, "Maternal depression and parenting behavior," *Clin. Psychol. Rev.*, vol. 20, no. 5, hal. 561–592, 2000, doi: 10.1016/s0272-7358(98)00100-7.
[25] M. R. Morales, S. Scherer, dan R. Levitan, "A cross-modal review of indicators for depression detection systems," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, hal. 1–12, 2017, doi: 10.18653/v1/w17-3101.
[26] M. B. Akçay dan K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, hal. 56–76, 2020, doi: 10.1016/j.specom.2019.12.001.
[27] S. Helmiyah, A. Fadlil, dan A. Yudhana, "Pengenalan Pola Emosi Manusia Berdasarkan Ucapan Menggunakan Ekstraksi Fitur Mel-Frequency Cepstral Coefficients (MFCC)," *CogITo Smart J.*, vol. 4, no. 2, hal. 372–381, 2019, doi: 10.31154/cogito.v4i2.129.372-381.
[28] Y. Shen, H. Yang, dan L. Lin, "Automatic Depression Detection: an Emotional Audio-Textual Corpus and a Gru/Bilstm-Based Model," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2022-May, hal. 6247–6251, 2022, doi: 10.1109/ICASSP43922.2022.9746569.
[29] A. Benba, A. Jilbab, dan A. Hammouch, "Analysis of multiple types of voice recordings in cepstral domain using MFCC for discriminating between patients with Parkinson's disease and healthy people," *Int. J. Speech Technol.*, vol. 19, no. 3, hal. 449–456, 2016, doi: 10.1007/s10772-016-9338-4.
[30] A. Sharif, O. S. Sitompul, dan E. B. Nababan, "Analysis Of Variation In The Number Of MFCC Features In Contrast To LSTM In The Classification Of English Accent Sounds," *J. Informatics Telecommun. Eng.*, vol. 6, no. 2, hal. 587–601, 2023, doi: 10.31289/jite.v6i2.8566.
[31] T. Jain, A. Jain, P. S. Hada, H. Kumar, V. K. Verma, dan A. Patni, "Machine Learning Techniques for Prediction of Mental Health," *Proc. 3rd Int. Conf. Inven. Res. Comput. Appl. ICIRCA 2021*, hal. 1606–1613, 2021, doi: 10.1109/ICIRCA51532.2021.9545061.
[32] W. W. Zung, "A Self-rating Depression Scale," *Arch. Gen. Psychiatry*, vol. 12, no. 1, hal. 63–70, 1965, doi: 10.1001/archpsyc.1965.01720310065008.
[33] J. Ancilin dan A. Milton, "Improved Speech Emotion Recognition with Mel Frequency Magnitude Coefficient," *Appl. Acoust.*, vol. 179, hal. 108046, 2021, doi: 10.1016/j.apacoust.2021.108046.
[34] V. Maheshwar, N. Venu Gopal, V. Naveen Kumar, D. Pranavi, dan Y. Padma Sai, "Development of an SVM-based Depression Detection Model using MFCC Feature Extraction," *Int. Conf. Sustain. Comput. Smart Syst. ICSCSS 2023 - Proc.*, no. Icscss, hal. 808–814, 2023, doi: 10.1109/ICSCSS57650.2023.10169770.
[35] M. J. Raj dan M. Sujith Kumar, "Gender based Affection Recognition of Speech Signals using Spectral & Prosodic Feature Extraction," *Int. J. Eng. Res. Gen. Sci.*, vol. 3, no. 2, hal. 898–905, 2015, [Daring]. Tersedia pada: www.ijergs.org
[36] A. Zlotnik, J. M. Montero, R. San-Segundo, dan A. Gallardo-Antolín, "Random forest-based prediction of Parkinson's disease progression using acoustic, ASR and intelligibility features," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2015-Janua, hal. 503–507, 2015, doi: 10.21437/interspeech.2015-184.

https://doi.org/10.33998/processor.2025.20.2.2513
Submitted: 13 Agustus 2025; Reviewed: 02 September 2025; Accepted; 11 September 2025; Published: 31 Oktober 2025

[37] A. Ahmed, R. Sultana, M. T. R. Ullas, M. Begom, M. M. I. Rahi, dan M. A. Alam, "A Machine Learning Approach to detect Depression and Anxiety using Supervised Learning," *2020 IEEE Asia-Pacific Conf. Comput. Sci. Data Eng. CSDE 2020*, 2020, doi: 10.1109/CSDE50874.2020.9411642.

[38] A. Arya, R. Kumari, dan P. Bansal, "Predicting Depression and Mental Illness Using Machine Learning Algorithms," *2023 Int. Conf. Commun. Secur. Artif. Intell. ICCSAI 2023*, hal. 399–404, 2023, doi: 10.1109/ICCSAI59793.2023.10421262.

[39] A. B. Abdusalomov, F. Safarov, M. Rakhimov, B. Turaev, dan T. K. Whangbo, "Improved Feature Parameter Extraction from Speech Signals Using Machine Learning Algorithm," *Sensors*, vol. 22, no. 21, 2022, doi: 10.3390/s22218122.

[40] A. Afshan, J. Guo, S. J. Park, V. Ravi, J. Flint, dan A. Alwan, "Effectiveness of voice quality features in detecting depression," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018-Septe, no. September, hal. 1676–1680, 2018, doi: 10.21437/Interspeech.2018-1399.

[41] M. Azhar dan H. F. Pardede, "Klasifikasi Dialek Pengujar Bahasa Inggris Menggunakan Random Forest," *J. Media Inform. Budidarma*, vol. 5, no. 2, hal. 439, 2021, doi: 10.30865/mib.v5i2.2754.

[42] N. Adhikari, S. Bhattacharya, M. Sultana, A. Maiti, dan D. Sengupta, "Identification of Bird via Acoustics for Bird Species Population Monitoring and Conservation," *Int. Conf. Big Data Anal. Bioinformatics, DABCon 2024*, hal. 1–6, 2024, doi: 10.1109/DABCon63472.2024.10919276.

https://doi.org/10.33998/processor.2025.20.2.2513
Submitted: 13 Agustus 2025; Reviewed: 02 September 2025; Accepted; 11 September 2025; Published: 31 Oktober 2025