Terakreditasi SINTA Peringkat 2
Surat Keputusan Direktur Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Nomor: 158/E/KPT/2021
masa berlaku mulai Volume 5 Nomor 2 Tahun 2021 sampai Volume 10 Nomor 1 Tahun 2026

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 6 No. 1 (2022) 115 - 122 ISSN Media Elektronik: 2580-0760

Pengenalan Emosi Pembicara Menggunakan *Convolutional Neural Networks*

Rendi Nurcahyo[1], Mohammad Iqbal[2]
[1]Teknik Elektro, Fakultas Teknik Elektro dan Teknologi Informasi, Universitas Gunadarma
[2]Ilmu Komputer, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Gunadarma
[1]rendi608@gmail.com, [2]mohiqbal@staff.gunadarma.ac.id

Abstract

Recognition of the speaker's emotions is an important but challenging component of Human-Computer Interaction (HCI). The need for the recognition of the speaker's emotions is also increasing related to the need for digitizing the company's operational processes related to the implementation of industry 4.0. The use of Deep Learning methods is currently increasing, especially for processing unstructured data such as data from voice signals. This study tries to apply the Deep Learning method to classify the speaker's emotions using an open dataset from SAVEE which contains seven classes of voice emotions in English. The dataset will be trained using the CNN model. The final accuracy of the model is 88% on the training data and 52% on the test data, which means the model is overfitting. This is due to the imbalance of emotion classes in the dataset, which makes the model tend to predict classes with more labels. In addition, the lack of heterogeneity of the dataset makes the character of the emotion class more different from the others so that it can reduce the bias in the model so as not to overfit the model. Further development of this research can be done, such as over-sampling the existing dataset by adding other data sources, then performing data augmentation to get the data character of each emotion class and setting hyperparameter values to get better accuracy values.

Keywords: Convolutional Neural Networks, Deep Learning, Keras, Speech Emotion Recognition, Tensorflow

Abstrak

Pengenalan emosi pembicara merupakan komponen penting namun menantang dari Interaksi Manusia-Komputer (HCI). Kebutuhan dari pengenalan emosi pembicara juga semakin meningkat terkait kebutuhan digitalisasi proses operasional perusahaan terkait implementasi industri 4.0. Penggunaan metode *Deep Learning* saat ini sudah semakin meningkat khususnya untuk pengolahan data tidak terstruktur seperti data dari sinyal suara. Penelitian ini mencoba mengimplementasikan metode *Deep Learning* untuk mengklasifikasikan emosi dari pembicara menggunakan dataset terbuka dari SAVEE yang berisi tujuh kelas emosi suara dalam bahasa Inggris. Dataset tersebut akan dilatih menggunakan model CNN. Akurasi akhir dari model sebesar 88% pada data latih dan 52% pada data tes yang berarti model mengalami *overfitting*. Hal tersebut disebabkan oleh tidak seimbangnya kelas emosi pada dataset sehingga membuat model akan cenderung memprediksi kelas yang labelnya lebih banyak. Selain itu, kurangnya heterogenitas dari dataset yang membuat karakter kelas emosi lebih berbeda dari yang lainnya sehingga mampu mengurangi bias pada model agar tidak membuat model *overfitting*. Pengembangan lebih lanjut dari penelitian ini yang dapat dilakukan seperti *over sampling* dataset yang dimiliki dengan menambahkan sumber data lain, melakukan augmentasi data agar mendapatkan karakter data setiap kelas emosi, dan mengatur nilai *hyperparameter* untuk mendapatkan nilai akurasi yang lebih baik.

Kata kunci: *Convolutional Neural Networks, Deep Learning,* Keras, *Speech Emotion Recognition, Tensorfl*

1. Pendahuluan

Pengenalan emosi pembicara adalah tindakan mencoba mengenali emosi manusia dan keadaan afektif dari ucapan. Ini memanfaatkan fakta bahwa suara sering kali mencerminkan emosi yang mendasari melalui nada dan nada. Ini juga merupakan fenomena yang digunakan hewan seperti anjing dan kuda untuk dapat memahami emosi manusia.

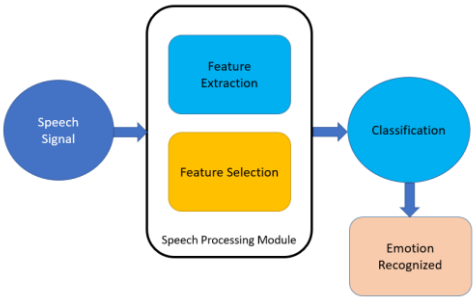
Pengenalan emosi dari ucapan telah berevolusi dari sesuatu yang khusus menjadi komponen penting untuk *Human-Computer Interaction* (HCI) [1] [3]. Sistem ini bertujuan untuk memfasilitasi interaksi alami dengan

Diterima Redaksi: 23-12-2021 | Selesai Revisi: 22-02-2022 | Diterbitkan Online: 27-02-2022

mesin melalui interaksi suara langsung alih-alih menggunakan perangkat tradisional sebagai masukan untuk memahami konten verbal dan memudahkan pendengar manusia untuk bereaksi [4] [6]. Beberapa aplikasi menyertakan sistem dialog untuk bahasa lisan seperti percakapan *call center*, *on board* sistem mengemudi kendaraan dan pemanfaatan pola emosi dari pidato dalam aplikasi medis [7]. Meskipun begitu, ada banyak masalah dalam sistem HCI yang masih perlu diperbaiki dan ditangani dengan benar, terutama karena sistem ini bergerak dari pengujian lab ke aplikasi dunia nyata [8] [10]. Oleh karena itu, upaya diperlukan untuk secara efektif memecahkan masalah tersebut dan mencapai pengenalan emosi yang lebih baik oleh mesin.

Menentukan keadaan emosional manusia adalah hal yang istimewa dan dapat digunakan sebagai standar untuk model pengenalan emosi apa pun [11]. Di antara banyak model yang digunakan untuk kategorisasi emosi ini, pendekatan emosional diskrit dianggap sebagai salah satu pendekatan fundamental. Ia menggunakan berbagai emosi seperti marah, bosan, jijik, kejutan, ketakutan, kegembiraan, kebahagiaan, netral dan kesedihan [12], [13]. Model penting lainnya yang digunakan adalah model tiga dimensi ruang kontinu dengan parameter seperti gairah, valensi, dan potensi.

Sistem pengenalan emosi berdasarkan pidato digital terdiri dari tiga komponen mendasar: sinyal preprocessing, ekstraksi fitur, dan klasifikasi [19]. Preprocessing akustik seperti denoising, serta segmentasi dilakukan untuk menentukan unit sinyal yang bermakna [20]. Ekstraksi fitur digunakan untuk mengidentifikasi fitur relevan yang tersedia dalam sinyal. Terakhir, pemetaan vektor fitur yang diekstraksi ke emosi yang relevan dilakukan oleh pengklasifikasi. Pada bagian ini, pembahasan rinci tentang pemrosesan sinyal suara, ekstraksi fitur, dan klasifikasi disediakan [21]. Juga, perbedaan antara ucapan spontan dan tindak tutur dibahas karena relevansinya dengan topik [22], [23]. Gambar 1 menggambarkan sistem sederhana yang digunakan untuk pengenalan emosi berbasis ucapan.



Gambar 1. Sistem pengenalan emosi suara tradisional

Pada tahap pertama pemrosesan sinyal berbasis ucapan, peningkatan suara dilakukan di mana komponen yang bising dihilangkan. Tahap kedua melibatkan dua bagian, ekstraksi fitur, dan seleksi fitur. Fitur yang diperlukan diekstraksi dari sinyal ucapan yang telah diproses sebelumnya dan pemilihan dibuat dari fitur yang diekstraksi. Ekstraksi dan pemilihan fitur tersebut biasanya didasarkan pada analisis sinyal suara dalam domain waktu dan frekuensi. Selama tahap ketiga, berbagai pengklasifikasi seperti GMM dan HMM, dll digunakan untuk klasifikasi fitur ini. Terakhir, berdasarkan klasifikasi fitur, emosi yang berbeda dikenali.

Suara manusia bersifat analog dan membawa sebuah informasi. Ada beberapa parameter yang dibawa oleh sinyal analog dari data audio yaitu seperti frekuensi, amplitudo, dan waktu. Parameter-parameter tersebut dapat dijadikan fitur untuk pembelajaran model *Deep Learning.*

Deep learning merupakan subbidang *Machine Learning* yang algoritmanya terinspirasi dari struktur otak manusia. Struktur tersebut dinamakan *Artificial Neural Networks* atau disingkat ANN. Pada dasarnya, ia merupakan jaringan saraf yang memiliki tiga atau lebih lapisan ANN. Ia mampu belajar dan beradaptasi terhadap sejumlah besar data serta menyelesaikan berbagai permasalahan yang sulit diselesaikan dengan algoritma *Machine Learning* lainnya.

Salah satu jenis algoritma *Deep Learning* yang paling populer dikenal sebagai *Convolutional Neural Network* (CNN atau ConvNet). CNN menggabungkan fitur yang dipelajari dengan data input, dan menggunakan lapisan konvolusi 2D, membuat arsitektur ini cocok untuk memproses data 2D, seperti gambar.

CNN menghilangkan kebutuhan untuk ekstraksi fitur manual, sehingga kita tidak perlu mengidentifikasi fitur yang digunakan untuk mengklasifikasikan gambar. CNN bekerja dengan mengekstraksi fitur langsung dari gambar. Fitur yang relevan tidak dilatih sebelumnya mereka dipelajari saat jaringan melatih kumpulan gambar. Ekstraksi fitur otomatis ini membuat model pembelajaran mendalam menjadi sangat akurat untuk tugas *Computer Vision* seperti klasifikasi objek.

Dalam studi ini, kami mencoba menggunakan metode *Deep Learning* dengan algoritma CNN (*Convolutional Neural Networks*) untuk mempelajari data sinyal suara agar mampu mengenali jenis emosi yang terdapat pada suara manusia.

2. Metode Penelitian

2.1. Pengumpulan Dataset dan Preprocessing Data

Pengumpulan data pada penelitian ini bersumber dari *Surrey Audio-Visual Expressed Emotion* (Savee) yang merupakan sumber terbuka untuk basis data suara yang dibuat oleh Universitas Surrey (United Kingdom) khusus untuk pengenalan emosi pada suara. Pada basis data tersebut terdapat 7 emosi suara dari 4 pria

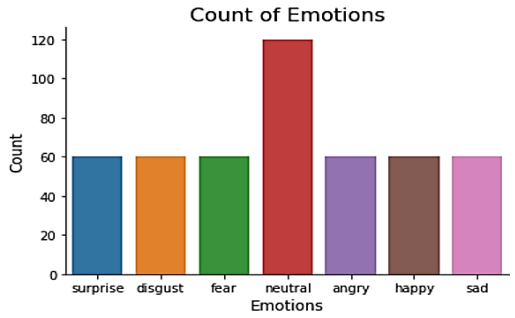
DOI: https://doi.org/10.29207/resti.v6i1.3726

menggunakan bahasa Inggris secara keseluruhan. Kalimat-kalimat yang digunakan pada basis data tersebut dipilih dari standar korpus TIMIT dan secara forentis seimbang untuk setiap emosi.

Tabel 1. Contoh isi dataset

Emotions	Path
surprise	data/DC_su09.wav
surprise	data/JK_su09.wav
disgust	data/DC_d10.wav
disgust	data/DC_d01.wav
fear	data/DC_f09.wav

Tabel 1 merupakan sampel dari dataset SAVEE yang akan dijadikan dataset untuk pelatihan model pengenalan emosi pembicara. Basis data SAVEE dipilih karena merupakan salah satu basis data terbuka yang menjadi tolak ukur untuk pembelajaran mesin dalam hal klasifikasi emosi. Basis data SAVEE telah teruji dan berhasil mendapatkan akurasi tinggi untuk klasifikasi emosi pada model pembelajaran mesin dengan menggunakan fitur sederhana. Pada penelitian ini kami akan mencoba basis data SAVEE untuk *Deep Learning* menggunakan algoritma CNN.

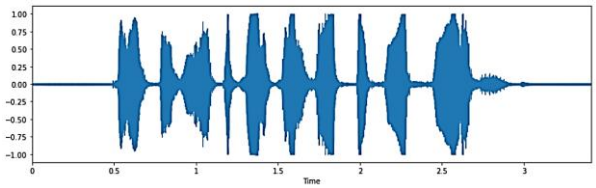


Gambar 2. Visualisasi dari dataset

Pada Gambar 2 terlihat jumlah label data untuk kelas "neutral" tidak seimbang, oleh karena itu dataset tersebut perlu dilakukan preprocessing data agar model tidak mengalami *overfitting* yang disebabkan oleh dataset yang tidak seimbang.

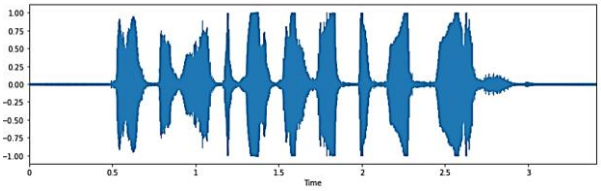
2.2. Augmentasi Data

Augmentasi data dalam analisis data adalah teknik yang digunakan untuk meningkatkan jumlah data dengan menambahkan salinan yang sedikit dimodifikasi dari data yang sudah ada atau data sintetis yang baru dibuat dari data yang ada. Pada penelitian ini augmentasi data audio menggunakan bantuan dari library Python yaitu librosa yang dapat melakukan pengolahan sinyal dari data audio.



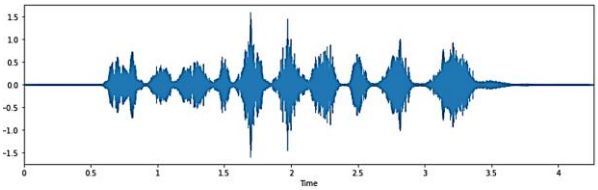
Gambar 3. Visualisasi gelombang salah satu sampel audio

Gambar 3 merupakan contoh visualisasi gelombang data audio normal dari dataset.



Gambar 4. Visualisasi gelombang salah satu sampel audio dengan *noise injection*

Gambar 4 merupakan contoh visualisasi gelombang data audio dengan menambahkan sedikit gangguan pada data.



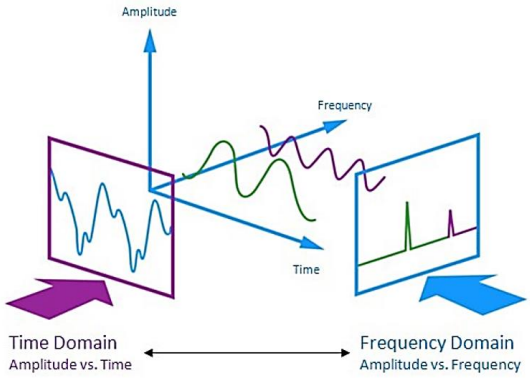
Gambar 5. Visualisasi gelombang salah satu sampel audio dengan dilakukan peregangan

Pada Gambar 5 merupakan contoh augmentasi data dengan dilakukan perenggangan pada gelombang data audio.

Dari gambar-gambar di atas merupakan contoh augmentasi data pada sampel dataset yang bertujuan membuat model menjadi invarian terhadap gangguan tersebut dan meningkatkan kemampuannya untuk menggeneralisasi.

2.3. Ekstraksi Fitur

Ekstraksi fitur adalah bagian yang sangat penting dalam menganalisis dan menemukan hubungan antara hal-hal yang berbeda. Data audio tidak dapat dipahami oleh model secara langsung sehingga perlu mengubahnya menjadi format yang dapat dimengerti untuk digunakan melalui ekstraksi fitur. Sinyal audio adalah sinyal tiga dimensi di mana tiga sumbu mewakili waktu, amplitudo dan frekuensi.



Gambar 6. Visualisasi fitur dari data audio

Pada penelitian ini kami tidak mendalami proses pemilihan fitur untuk memeriksa fitur mana yang bagus

DOI: https://doi.org/10.29207/resti.v6i1.3726

untuk dataset kami, melainkan hanya mengekstrak 5 fitur berikut beserta fungsinya: (1). *Zero Crossing Rate*: Tingkat perubahan tanda sinyal selama durasi frame tertentu. Digunakan untuk klasifikasi karakter suara antar *frame*, (2). *Chroma_stft*: Chroma STFT Nilai Chroma dari suatu audio pada dasarnya mewakili intensitas dua belas kelas nada khusus yang digunakan untuk mempelajari musik. Mereka dapat digunakan dalam diferensiasi profil kelas nada antara sinyal audio, (3). MFCC (*Mel Frequency Cepstral Coefficients*): Fungsi MFCC memproses seluruh data ucapan dalam satu batch. Berdasarkan jumlah baris input, panjang jendela, dan panjang tumpang tindih, MFCC mempartisi suara menjadi 1551 frame dan menghitung fitur Cepstral untuk setiap frame. Setiap baris dalam matriks koefisien sesuai dengan nilai log-energi diikuti oleh koefisien Cepstral frekuensi 13 mel untuk frame yang sesuai dari file suara. Fungsi ini juga menghitung lokasi sampel terakhir di setiap frame input, (4). RMS (*Root Mean Square*) value: Sebagai metriks dari efektivitas model, (5). Mel Spectogram: Spektogram Mel digunakan untuk menyediakan model dengan informasi suara yang mirip dengan apa yang akan dirasakan manusia. Bentuk gelombang audio mentah dilewatkan melalui bank filter untuk mendapatkan spektogram Mel. Setelah proses ini, setiap sampel memiliki bentuk 128 x 128, yang menunjukkan 128 bank filter yang digunakan dan 128 langkah waktu per klip.

Nilai dari fitur-fitur tersebut didapatkan dari file audio yang dijadikan dataset lalu diekstrak menggunakan bantuan pustaka pada Python yaitu Librosa. Terdapat 163 kolom data nilai dari fitur yang telah diekstrak untuk setiap file audio dan dikumpulkan dalam 1 dataset yang nantinya akan digunakan untuk pelatihan model.

	0	1	2	3	4	5	6	7	8	9	...
0	0.021828	0.526627	0.428081	0.374749	0.395473	0.408206	0.435075	0.488643	0.522054	0.614687	...
1	0.046880	0.633381	0.578276	0.549857	0.558332	0.570838	0.549701	0.557231	0.597113	0.661779	...
2	0.023362	0.521085	0.514815	0.393702	0.341573	0.354879	0.367003	0.376753	0.415249	0.478445	...
3	0.073636	0.527316	0.535347	0.599175	0.620190	0.546638	0.549203	0.520936	0.452054	0.474522	...
4	0.094541	0.561282	0.572799	0.642505	0.676083	0.606296	0.569471	0.526151	0.471762	0.490605	...

5 rows × 163 columns

Gambar 7. Sampel data *input* model yang akan digunakan pada saat pelatihan

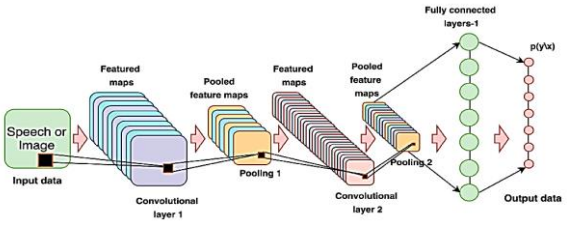
2.4. Pelatihan Model dan Evaluasi

Penelitian ini menggunakan salah satu model *Deep Learning* untuk klasifikasi yaitu *Convolutional Neural Networks* (CNN) pada Keras yang dijalankan diatas Tensorflow.

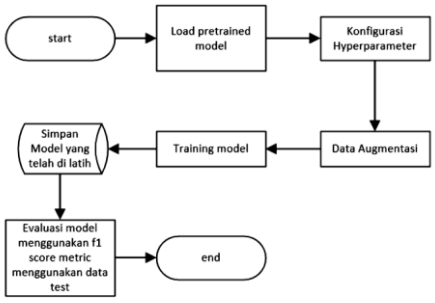
CNN adalah jenis lain dari teknik *Deep Learning* yang hanya didasarkan pada arsitektur *feed forward* [24] untuk klasifikasi. CNN biasanya digunakan untuk pengenalan pola dan memberikan klasifikasi data yang lebih baik. Jaringan ini memiliki neuron ukuran kecil yang ada pada setiap lapisan arsitektur model yang dirancang yang memproses data input dalam bentuk bidang reseptif [25]. Gambar 8 menunjukan arsitektur *layer-wise* dari jaringan CNN dasar. Filter adalah basis koneksi lokal yang dililitkan dengan input dan berbagi parameter yang sama (bobot Wi dan bias ni) untuk menghasilkan i peta fitur (zi), masing-masing berukuran a – b – 1. Lapisan konvolusi menghitung produk titik antara bobot dan input yang diberikan. Jadi, parameter untuk bobot W^i dan bias n^i untuk pembuatan peta z^i untuk fitur i dengan ukuran a – b – 1 dapat diberikan sebagai

$$z^i = g(W^i * r + n^i) \quad (1)$$

Fungsi aktivasi f atau metodologi non-linier perlu diterapkan untuk mendapatkan output dari lapisan konvolusi. Perlu dicatat bahwa input adalah daerah yang sangat kecil dari volume asli seperti yang digambarkan pada Gambar 10. Pengambilan sampel bawah dilakukan pada setiap lapisan subsampling untuk menampilkan peta dan mengurangi parameter dalam jaringan. Ini, pada gilirannya, mengontrol *overfitting* dan meningkatkan proses pelatihan. Proses penyatuan dilakukan pada elemen pp (juga dikenal sebagai ukuran filter) untuk bentangan semua peta fitur yang berdampingan. Pada tahap akhir, lapisan harus terhubung sepenuhnya seperti pada jaringan saraf lainnya. Lapisan selanjutnya ini mengambil fitur tingkat rendah dan tingkat menengah sebelumnya dan menghasilkan abstraksi tingkat tinggi dari data ucapan masukan. Lapisan terakhir juga dikenal sebagai SVM atau Softmax digunakan untuk lebih menghasilkan skor klasifikasi dalam istilah probabilistik untuk berhubungan dengan kelas tertentu.



Gambar 8. Arsitektur *layer-wise* jaringan dasar CNN



Gambar 9. Alur penelitian

DOI: https://doi.org/10.29207/resti.v6i1.3726

Proses latih dan proses evaluasi pada penelitian ini dilakukan untuk mengetahui performance dari model terhadap hyperparameter. Hyperparameter yang digunakan pada penelitian ini berupa learning rate, momentum, epoch dan fully connected layer selain itu penelitian ini menggunakan Data Augmentation saat proses latih berlangsung. Variabel data Augmentation yang digunakan adalah zooming, shear_range dengan nilai 0.2 yang dimana nilai 0.2 tersebut merupakan nilai persenan dari citra tersebut, jika nilai zooming 0.2 maka citra akan di zooming 0.2 % dari ukuran citra asli selain itu flipping image juga digunakan pada Data Augmentation pada proses evaluasi performa dari model dilakukan setelah suatu model telah dilatih, proses evaluasi model pada penelitian ini menggunakan data test yang telah dipersiapkan ini dilakukan untuk mengetahui performa dari model tertentu terhadap citra di luar dataset latih, selain itu metric yang digunakan pada penelitian ini adalah precision, recall, dan juga *f1 score* dengan rumus 2, 3, dan 4.

$$Precision\ (\%) = \frac{TP}{(FP+TP)} \quad (2)$$

$$Recall\ (\%) = \frac{TP}{(FN+TP)} \quad (3)$$

$$F - Measure\ (\%) = \frac{(2\ x\ Recall\ x\ Precision)}{(Recall + Precision)} \quad (4)$$

Perhitungan F1 score, precision dan recall berdasarkan data TP, FN, FP dan TP yang didapatkan dari tabel confussion matrix. True Positive (TP) merupakan jumlah prediksi yang benar di kelas positif, sedangkan False Negative (FN) memiliki makna jumlah prediksi yang salah di kelas negatif, False Positive (FP) adalah jumlah prediksi yang salah di kelas positif sedangkan untuk True Negative (TN) memiliki arti jumlah prediksi yang benar di kelas negatif. Ilustrasi tabel confussion matrix dapat dilihat pada Gambar 10.

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Gambar 10. Ilustrasi confusion matrix dari prediksi

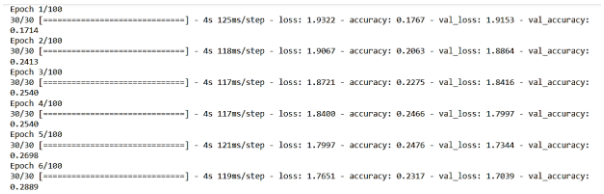
3. Hasil dan Pembahasan

Dengan menggunakan konfigurasi *Neural Network* sebagai berikut, tabel 2.

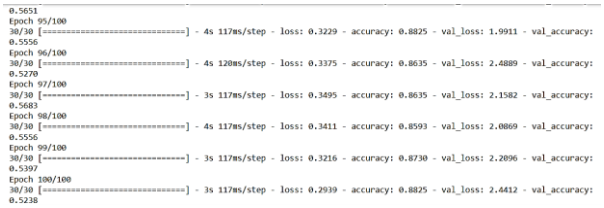
Hasil akhir dari pelatihan model klasifikasi emosi pembicara dengan CNN adalah 88% akurasi dengan nilai kehilangan 0,29 menggunakan data latih dan 52% akurasi dengan nilai kehilangan 2,44 menggunakan data tes untuk 100 periode latihan (dilihat pada metrik validasi akurasi).

Tabel 2. Konfigurasi *Neural Network*

Tipe Lapisan	Bentuk Keluaran	Parameter
conv1d (Conv1D)	(None, 162, 256)	1536
max_pooling1d (MaxPooling1D)	(None, 81, 256)	0
conv1d_1 (Conv1D)	(None, 81, 256)	327936
max_pooling1d_1 (MaxPooling1D)	(None, 41, 256)	0
conv1d_2 (Conv1D)	(None, 41, 128)	163968
max_pooling1d_2 (MaxPooling1D)	(None, 21, 128)	0
dropout (Dropout)	(None, 21, 128)	0
conv1d_3 (Conv1D)	(None, 21, 64)	41024
max_pooling1d_3 (MaxPooling1D)	(None, 11, 64)	0
flatten (Flatten)	(None, 704)	0
dense (Dense)	(None, 32)	22560
dropout_1 (Dropout)	(None, 32)	0
dense_1 (Dense)	(None, 7)	231

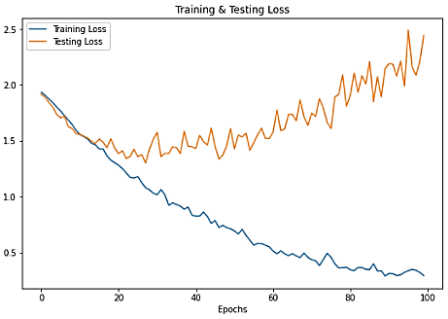


Gambar 11. Proses latih model iterasi awal



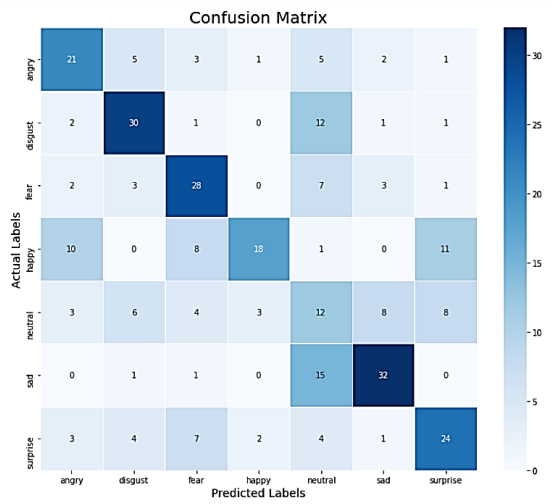
Gambar 12. Proses latih model iterasi akhir

Gambar 13. Grafik akurasi model



Gambar 14. Grafik kehilangan model

DOI: https://doi.org/10.29207/resti.v6i1.3726



Gambar 15. Confusion matrix dari prediksi

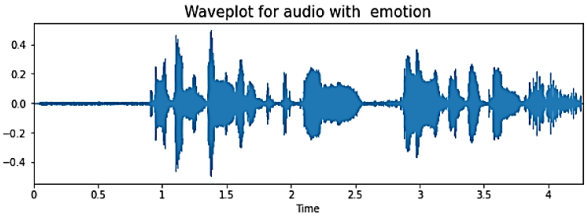
Tabel 3. Hasil prediksi masing-masing kelas dengan data tes

Kelas	Presisi	Balikan	F1-Score
Angry	51%	55%	53%
Disgust	61%	64%	62%
Fear	54%	64%	58%
Happy	75%	38%	50%
Neutral	21%	27%	24%
Sad	68%	65%	67%
Surprise	52%	53%	53%

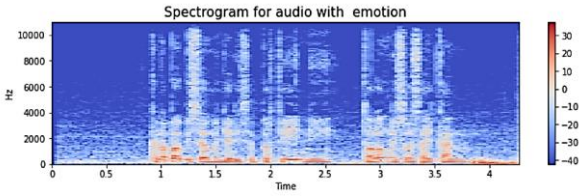
Berdasarkan hasil prediksi model dengan data tes pada tabel di atas dapat dilihat bahwa model lebih akurat dalam memprediksi disgust, sad, dan surprise. Hal tersebut dikarenakan file audio emosi ini berbeda dengan file audio lainnya dalam banyak hal seperti nada, kecepatan, dan tekanan intonasi.

Selain itu, alasan rendahnya akurasi dari model ketika menggunakan data tes dikarenakan model yang mengalami *overfitting* yang disebabkan beberapa hal seperti kurangnya heterogenitas dari data latih sehingga membuat model cenderung menghafal kondisi data latih yang menyebabkan model tidak dapat memprediksi data tes baru yang berbeda kondisi dengan data latih.

Pada penelitian ini dataset yang digunakan untuk data latih menggunakan bahasa Inggris. Kami ingin membuktikan bahwa model tersebut tidak tergantung pada bahasa yang digunakan karena fitur yang dibaca model adalah karakter sinyal analog dari suara untuk mendapatkan jenis intonasinya.



Gambar 16. Visualisasi gelombang audio tes bahasa Indonesia



Gambar 17. Visualisasi spektogram audio tes bahasa Indonesia

```
pred_new_data = model.predict(feature_new_data_dims)
pred_new_data_inverse = encoder.inverse_transform(pred_new_data)
pred_new_data_inverse.flatten()[0]
```

```
'angry'
```

Gambar 18. Hasil prediksi data tes dengan Bahasa Indonesia

Setelah dicoba dengan audio menggunakan bahasa Indonesia hasilnya model dapat memprediksi dengan tepat untuk kelas Angry karena memang kelas tersebut cukup baik F1-*score*-nya.

4. Kesimpulan

Dari hasil yang diperoleh dapat disimpulkan bahwa model *Convolutional Neural Network* (CNN) dengan metode *Deep Learning* setelah mempelajari dataset SAVEE mampu mengenali emosi dari suara namun belum cukup baik dengan akurasi model secara keseluruhan pada data tes masih sangat rendah yaitu 52% dengan konfigurasi *Hyperparameter* fungsi aktivasi yang digunakan yaitu 4 layer ReLU dan 1 layer Softmax dengan *optimizer* Adam, parameter *loss* menggunakan *categorical_crossentropy*, dan metriks acuan yaitu akurasi yang dilatih sebanyak 100 *epoch*.

F1-score masing-masing kelas dengan data tes yaitu *Angry* 53%, *Disgust* 62%, *Fear* 58%, *Happy* 50%, *Neutral* 24%, *Sad* 67%, *Surprise* 53% masih terlalu rendah. Hal tersebut disebabkan oleh tidak seimbangnya kelas emosi pada dataset sehingga membuat model akan cenderung memprediksi kelas yang labelnya lebih banyak. Selain itu, kurangnya heterogenitas dari dataset yang membuat karakter kelas emosi lebih berbeda dari yang lainnya sehingga mampu mengurangi bias pada model agar tidak membuat model *overfitting*.

Penelitian ini dapat dikembangkan lebih lanjut untuk meningkatkan akurasi dari model sehingga model tersebut dapat diaplikasikan untuk membuat sistem pengenalan emosi yang sangat berguna dan mampu meyelesaikan banyak masalah terutama untuk kebutuhan industri. Pengembangan yang dapat dilakukan seperti *over sampling* dataset yang dimiliki dengan menambahkan sumber data lain, melakukan augmentasi data agar mendapatkan karakter data setiap kelas emosi, dan mengatur nilai *hyperparameter* agar model dapat belajar lebih baik disetiap iterasi pelatihan.

Aplikasi dari pengenalan emosi ini juga sangat bermanfaat nantinya seperti digunakan pada *call center* untuk mengklasifikasikan panggilan menurut emosi dan

DOI: https://doi.org/10.29207/resti.v6i1.3726

dapat digunakan sebagai parameter kinerja untuk analisis percakapan sehingga mengidentifikasi pelanggan yang tidak puas, kepuasan pelanggan, dan sebagainya untuk membantu perusahaan meningkatkan layanan mereka. Dapat juga digunakan sistem *in-car board* berdasarkan informasi kondisi mental pengemudi yang dapat diberikan kepada sistem untuk menginisiasi keselamatannya mencegah terjadinya kecelakaan.

Daftar Rujukan

[1] B. W. Schuller, ``Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," Commun. ACM, vol. 61, no. 5, pp. 90-99, 2018.
[2] M. S. Hossain and G. Muhammad, ``Emotion recognition using deep learning approach from audio-visual emotional big data," Inf. Fusion, vol. 49, pp. 69-78, Sep. 2019.
[3] M. Chen, P. Zhou, and G. Fortino, ``Emotion communication system," IEEE Access, vol. 5, pp. 326-337, 2016.
[4] N. D. Lane and P. Georgiev, ``Can deep learning revolutionize mobile sensing?" in Proc. ACM 16th Int. Workshop Mobile Comput. Syst. Appl., 2015, pp. 117-122.
[5] J. G. Rázuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet, and A. Larsson, ``Speech emotion recognition in emotional feedback for human-robot interaction," Int. J. Adv. Res. Artif. Intell., vol. 4, no. 2, pp. 20-27, 2015.
[6] D. Le and E. M. Provost, ``Emotion recognition from spontaneous speech using hidden MARKOV models with deep belief networks," in Proc. IEEE Workshop Autom. Speech Recognit. Understand., Dec. 2013, pp. 216-221.
[7] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, ``Speech recognition using deep neural networks: A systematic review," IEEE Access, vol. 7, pp. 19143-19165, 2019.
[8] S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh, ``Speech emotion recognition," in Proc. Int. Conf. Adv. Electron. Comput. Commun. (ICAECC), Oct. 2014, pp. 1-4.
[9] K. R. Scherer, ``What are emotions? And how can they be measured?" Social Sci. Inf., vol. 44, no. 4, pp. 695-729, 2005.
[10] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias, ``Emotion analysis in man-machine interaction systems," in Proc. Int.Workshop Mach. Learn. Multimodal Interact. Springer, 2004, pp. 318-328.
[11] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, ``Emotion recognition in human computer interaction," IEEE Signal Process. Mag., vol. 18, no. 1, pp. 32-80, Jan. 2001.
[12] O. Kwon, K. Chan, J. Hao, T. Lee, ``Emotion recognition by speech signal," in Proc. EUROSPEECH, Geneva, Switzerland, 2003, pp. 125-128.
[13] R. W. Picard, ``Affective computing," Perceptual Comput. Sect., Media Lab., MIT, Cambridge, MA, USA, Tech. Rep., 1995.
[14] S. G. Koolagudi and K. S. Rao, ``Emotion recognition from speech: A review," Int. J. speech Technol., vol. 15, no. 2, pp. 99-117, 2012.
[15] M. El Ayadi, M. S. Kamel, and F. Karray, ``Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognit., vol. 44, no. 3, pp. 572-587, 2011.
[16] A. D. Dileep and C. C. Sekhar, ``GMM-based intermediate matching kernel for classi-cation of varying length patterns of long duration speech using support vector machines," IEEE Trans. neural Netw. Learn. Syst., vol. 25, no. 8, pp. 1421-1432, Aug. 2014.
[17] L. Deng and D. Yu, ``Deep learning: Methods and applications," Found. Trends Signal Process., vol. 7, nos. 3-4, pp. 197-387, Jun. 2014.
[18] J. Schmidhuber, ``Deep learning in neural networks: An overview," Neural Netw., vol. 61, pp. 85-117, Jan. 2015.
[19] T. Vogt and E. André, ``Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2005, pp. 474-477.
[20] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, ``Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011,"Artif. Intell. Rev., vol. 43, no. 2, pp. 155177, 2015.
[21] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, ``The automatic recognition of emotions in speech," in Emotion-Oriented Systems. Springer, 2011, pp. 71-99.
[22] E. Mower, M. J. Mataric, and S. Narayanan, ``A framework for automatic human emotion classication using emotion proles," IEEE Trans. Audio, Speech, Language Process., vol. 19, no. 5, pp. 1057-1070, Jul. 2011.
[23] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, ``Prediction-based learning for continuous emotion recognition in speech," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 5005-5009.
[24] W. Q. Zheng, J. S. Yu, and Y. X. Zou, ``An experimental study of speech emotion recognition based on deep convolutional neural networks," in Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII), Sep. 2015, pp. 827-831.
[25] F. Dipl and T. Vogt, "Real-time automatic emotion recognition from speech," 2010.
[26] S. Lugovic, I. Dunder, and M. Horvat, "Techniques and applications of emotion recognition in speech," 2016 39th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2016 - Proc., no. November 2017, pp. 1278–1283, 2016.
[27] F. Noroozi, N. Akrami, and G. Anbarjafari, "Speech-based emotion recognition and next reaction prediction," 2017 25th Signal Process. Commun. Appl. Conf. SIU 2017, no. 1, 2017.
[28] C.-W. Huang and S. S. Narayanan, "Characterizing Types of Convolution in Deep Convolutional Recurrent Neural Networks for Robust Speech Emotion Recognition," pp. 1–19, 2017.
[29] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.
[30] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," *2017 Int. Conf. Platf. Technol. Serv.*, pp. 1–5, 2017.

DOI: https://doi.org/10.29207/resti.v6i1.3726