Jurnal Informatika Universitas Pamulang **ISSN: 2541-1004**
Penerbit: Fakultas Ilmu Komputer Universitas Pamulang **e-ISSN: 2622-4615**
Vol. 9, No. 4, Desember 2024 (163-171) **https://doi.org/10.32493/informatika.v9i4.45236**

Klasifikasi Emosi Berdasarkan Suara dengan Metode *Convolutional Neural Network*

Muhammad Elio Phillo Rismanto[1*], Irma Handayani[2]

[12]Teknik Informatika, Universitas Teknologi Yogyakarta, Jl. Siliwangi (Ringroad Utara), Jombor, Sleman, D.I. Yogyakarta, Indonesia, 55285
e-mail: [1]muhammad.5210411291@student.uty.ac.id, [2]irma.handayani@staff.uty.ac.id
*Coressponding Author

Submitted Date: December 02, 2024 Reviewed Date: December 23, 2024
Revised Date: December 27, 2024 Accepted Date: December 28, 2024

Abstract

Voice-based emotion detection technology (SER), is the study of machines' ability to comprehend patterns in voice data, utilizing a range of methods and features. However, its utilizations remains limited due to the inherent challenges faced by machines in accurately discerning emotions. This research was conducted using a frequently used method, namely CNN and was developed to produce a high-accuracy method, with spectrogram features due to their capacity to record frequencies in RAVDESS. The data set comprised 2068 voice samples classified into five emotion classes: angry, afraid, happy, sad, and neutral. The augmentation of all data regarding noise, pitch, shifting, stretching, and high and low speed, was implemented to replicate real-world conditions. This research was conducted by training on several parameters such as: learning rate, dropout rate, kernel, weight decay size, optimization, epochs, and batch size. This research resulted in a CNN method with the best parameter values produced {weight_decay': 1e-07, 'optimizer': 'adamw', 'learning_rate': 0.001, 'kernel_initializer': 'he_normal', 'dropout_rate': 0.5, 'epochs': 100, 'batch_size': 48}, which has score value of 0.7448840381991815. The model demonstrated a general accuracy level of 75.85% for the training data and 51.64% for the test data, indicating its ability to recognize existing patterns but difficulty in generalizing new data. However, the ROC curve values indicate that the model is capable of differentiating voice data into its respective classes, with values of 0.84 for angry emotions, 0.79 for fear emotions, 0.83 for happy emotions, 0.80 for sad emotions, and 0.9 for neutral emotions.

Keywords: CNN; SER; RAVDESS

Abstrak

Teknologi deteksi emosi berdasarkan suara (SER) merupakan studi mengenai proses mesin untuk memahami pola dalam data suara yang memiliki beragam metode serta fitur. Teknologi deteksi ini penggunaannya masih sedikit karena susahnya mesin dalam memahami emosi. Penelitian ini dilakukan dengan menggunakan metode CNN lalu dikembangkan sehingga dihasilkan metode berakurasi tinggi. Metode CNN menggunak fitur spectrogram karena kemampuannya dalam merekam frekuensi RAVDESS yang berjumlah 2068 data suara dengan 5 kelas emosi yaitu marah, takut, senang, sedih dan netral. Augmentasi dilakukan terhadap seluruh data mengenai kebisingan suara, frekuensi, pergeseran, perenggangan, dan tinggi rendahnya kecepatan suara untuk mereplika kondisi di dunia nyata. Penelitian dilakukan dengan melatih parameter yang ditentukan seperti: tingkat pembelajaran, tingkat pemutusan, jenis kernel, bobot peluruhan, metode optimasi, jumlah *epoch* dan *batch*. Penelitian ini menghasilkan metode CNN dengan nilai terbaik {'bobot_peluruhan': 1e-07, 'optimasi': 'adamw', 'tingkat_pembelajaran': 0.001, 'kernel_inisialisasi': 'he_normal', 'tingkat_pemutusan': 0.5, 'epochs': 100, 'ukuran_batch': 48}, dan hasil skor 0.7448840381991815 serta tingkat akurasi umum dihasilkan data latih sebesar 75.85% dan data uji sebesar 51.64% yang menandakan model menghafal pola namun kesusahan untuk melakukan generalisasi data baru. Walaupun demikian, jika melihat pada nilai kurva ROC, emosi marah di prediksi

0.84, emosi takut 0.79, emosi senang 0.83, emosi sedih 0.80 dan emosi netral 0.9. Nilai kurva ini menandakan bahwa model cukup mampu untuk membedakan data suara ke kelasnya masing–masing.

Kata Kunci: CNN; SER; RAVDESS

1. Pendahuluan

Teknologi deteksi emosi berdasarkan suara atau yang lebih dikenal sebagai *Speech Emotion Recognition* (SER) merupakan salah satu studi yang mempelajari bagaimana mesin memahami pola yang ada dari perkataan atau suara manusia (George & Muhamed Ilyas, 2024). Kemudian, Pola – pola yang muncul di dalam perkataan maupun suara manusia ini akan diasumsikan sebagai parameter-parameter tertentu yang ciri khasnya mampu digeneralisasi menjadi suatu kelas emosi (Juslin & Scherer, 2008).

Kemampuan Teknologi SER dalam mengenali pola dari emosi dalam suara manusia dapat diaplikasikan ke dalam berbagai bidang pengetahuan dan memberikan manfaat yang besar. Walaupun demikian, Teknologi SER masih jarang digunakan di khalayak umum. Hal ini karena sulitnya mesin dalam memahami emosi manusia terutama dari suara. Suara manusia sangat beragam baik dari intonasi, keras kecilnya suara, logat, bahkan gender dapat menjadi faktor penentu emosi.

Penelitian ini menghasilkan metode SER yang mampu memprediksi emosi manusia dari suaranya dengan baik. Hal ini dilakukan dengan menggunakan salah satu metode yang sudah matang di dalam SER yaitu metode *Convolutional Neural Network* (CNN). Metode CNN ini termasuk ke dalam bagian dari *computer vision* yang bekerja dengan mengekstraksi fitur–fitur yang ada dan kemudian melalui lapisan–lapisan seperti konvolusi, aktivasi, dan pooling dihasilkan klasifikasi (Khan et al., 2020). Dengan kata lain, metode CNN ini mampu mendeteksi gambar, video, dan suara. Metode CNN dipilih selain kematangannya dalam SER juga karena kemampuannya dalam mengatasi data dengan jumlah besar dan rumit (Bhatt et al., 2021), bahkan dengan daya komputasi yang terbatas (Salehi et al., 2023).

Penelitian relevan pertama dengan penelitian ini dilakukan oleh Aini et al. Aini et al (2021) melakukan perbandingan data suara yang didapat melalui fitur *Mel-Frequency Cepstral Coeffisients* (MFCC) frekuensi fundamental dengan *Root Mean Square Energy* (RMSE) yang selanjutnya dilakukan klasifikasi melalui metode CNN. Klasifikasi digunakan berupa netral, sedih, senang dan marah. Penelitian tersebut menghasilkan tingkat akurasi sebesar 85% dengan berdasarkan gabungan fitur MFCC dan Frekuensi fundamental serta akurasi sebesar 72 % dengan gabungan antara fitur MFCC dengan RMSE. Penelitian tersebut juga menghasilkan kesimpulan bahwa fitur RMSE tidak cocok untuk digunakan dalam deteksi emosi berbasis suara.

Penelitian relevan kedua dilakukan oleh Tanudjaja et.al. Menurut Tanudjaja et.al (2023) deteksi emosi dengan menggunakan data The Ryerson Audio-Visual Dataset of Emotional Speech and Song (RAVDESS) didapat melalui laman Kaggle. Fitur – fitur yang digunakan antara lain MFCC, Energi, Pitch, Spectral Centroid, Spectral Flatness, dan Spectral Roll-off. Penelitian tersebut melakukan pengujian terhadap emosi tenang, netral, sedih, senang, takut, marah, terkejut, dan jijik. Dan menghasilkan nilai akurasi sebesar 70% berdasarkan fitur MFCC dengan rata – rata prediksi benar sebesar 77 % pada emosi marah, terkejut, sedih dan tenang.

Penelitian relevan ke tiga dilakukan oleh Aluhaidan. Dalam penelitiannya, Alluhaidan et al. (2023) menggunakan data set Emo–DB, SAVEE, dan RAVDESS juga menghasilkan tingkat akurasi sebesar 97% pada Emo – DB data set, 93% pada SAVEE data set, dan 92 % pada RAVDESS data set. Penelitian relevan ke empat dilakukan oleh Rahmadani et al. (2022) dengan metode *deep learning* CNN seperti BiLSTM, BiGRU, CNN–BiLSTM, dan CNN–BiGRU menghasilkan prediksi akurasi tertinggi pada 91,29% dengan metode CNN–BiLSTM sementara akurasi terendah sebesar 81,34% dengan metode LSTM.

Dengan metode CNN sebagai dasarnya, perlu dilakukan pemilihan fitur yang akan digunakan dalam penelitian ini. Spectrogram dipilih atas kemampuannya untuk merekam frekuensi baik tinggi maupun rendahnya dari suatu suara (Gencyilmaz & Karaoğlan, 2024; Li et al., 2022). Dengan demikian, pergeseran frekuensi yang dihasilkan spectrogram ini akan membantu CNN untuk mengenali pola lebih mudah dalam suatu kelas emosi.

Fitur spectrogram banyak digunakan dalam penelitian, tetapi penggunaannya jarang pada klasifikasi emosi. Salah satu penelitian fitur

spectrogram untuk mengklasifikasikan jenis burung berdasarkan perbedaan kicauannya dilakukan oleh Hu et al. Penelitian Hu et al. (2023) yang dilakukan terhadap tiga data set yaitu data set Huabei, data set Urbansound8K dan data set Birdsdata menghasilkan tingkat akurasi sebesar 96,28% untuk data set Huabei, 98,34% untuk data set Urbansound8K dan akurasi sebesar 96.66% untuk dataset Birdsdata. .

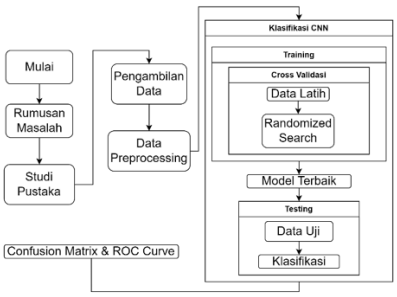
Penelitian lain yang menggunakan fitur spectrogram dalam bidang medis dilakukan oleh Ozcelik et al. Ozcelik et al (2023) melakukan penelitian dengan memetakan spectrogram untuk mendeteksi tingkat tekanan darah dan melakukan klasifikasi apakah tekanan darah seorang pasien normal atau terjadi hipertensi. Hasil penelitian ini menghasilkan tingkat akurasi dengan model usulan sebesar 98,14%, 98,79%, dan 97,69% untuk arsitektur ResNet18, ResNet50, dan ConvMixer, masing-masing.

Berdasarkan uraian di atas, deteksi emosi fitur MFCC merupakan fitur yang sering digunakan dan menghasilkan akurasi rentang 70 % hingga 97% dengan berbagai macam metode dan kombinasinya yang ada. Walaupun fitur MFCC mampu dengan baik memprediksi emosi dan menjadi fitur utama banyak penelitian, tetapi hal ini menjadi alasan peneliti untuk menggunakan fitur spectrogram sebagai fitur yang dipilih karena sedikitnya penelitian penggunaan fitur ini dalam mendeteksi emosi berdasarkan suara. Selain itu, penggunaan fitur spectrogram dan metode CNN diharapkan mampu menghasilkan prediksi emosi suara manusia yang lebih baik dari model di penelitian relevan yang dikemukakan yang mana di penelitian ini, parameter- parameter dalam pembentukan model akan di ujikan untuk menghasilkan model terbaik.

2. Metode Penelitian

2.1. Tahapan Penelitian

Penelitian ini memiliki tahapan yang digunakan sebagai gambar 1 berikut..



Gambar 1 Metode Penelitian

2.2. Rumusan Masalah

Pada Penelitian ini, permasalahan yang dibahas adalah bagaimana mendesain Metode CNN yang mampu mengatasi data suara berupa Spectrogram dengan berbagai kelas emosi

2.3. Studi Pustaka

Studi - studi pustaka yang digunakan diperoleh melalui jurnal – jurnal berkaitan dengan metode CNN, Spectrogram dan Deteksi Emosi sebagai landasan penelitian ini

2.4. Pengambilan Data

Data yang digunakan diperoleh melalui Kaggle dan berupa data audio RAVDESS. Dalam data tersebut terdapat 24 aktor dengan 8 jenis suara, yaitu tenang, bahagia, sedih, marah, takut, terkejut, dan jijik (Livingstone & Russo, 2018). Pada Penelitian ini, kelas emosi yang digunakan hanyalah 5 kelas yaitu marah, takut, senang, sedih dan neutral dengan total data sebesar 2068 data.

2.5. Pre Processing

Penelitian ini menggunakan beberapa tahapan preprocessing. Tahapan pertama adalah preprocessing dengan melakukan ekstraksi fitur Spectrogram dan kemudian dilakukan augmentasi data dengan melakukan

1. Noise : Dilakukan untuk menghasilkan suara acak untuk menyimulasi kebisingan
2. Pitch : Mengubah tinggi rendahnya nada di dalam data audio
3. Scretch : Mengubah kecepatan audio tanpa mengubah nada di dalam data
4. Shift : Menggeser Data audio secara acak ke kiri maupun ke kanan
5. Higher speed : Meningkatkan Kecepatan Audio
6. Lower speed : Menurunkan kecepatan audio

Augmentasi ini dilakukan untuk memberikan variasi data sehingga metode CNN dapat belajar untuk mengatasi data suara yang kurang baik (Mumuni & Mumuni, 2022)

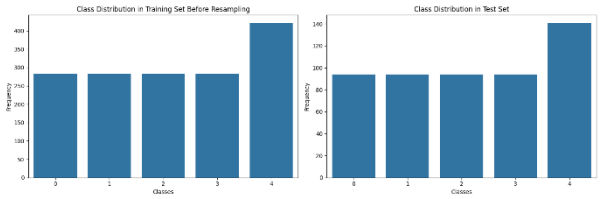
Tahapan selanjutnya adalah dengan dilakukannya padding terhadap data hasil ekstraksi Spectrogram sehingga mampu untuk dilakukan normalisasi dan oversampling. Tahapan Padding dilakukan dengan melihat panjang maksimum dari Spectrogram dan dilakukan pengisian dengan nilai 0. Selanjutnya data dapat diolah melalui Normalisasi *Min Max Scaler* dan algoritma *ADASYN*.

Normalisasi Min Max Scaler dilakukan untuk mengubah atau menormalkan data

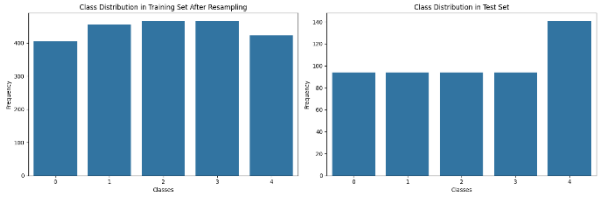
fitur spectrogram yang bervariasi menjadi di rentang antara nol dan satu (Deepa & Ramesh, 2022) Setelahnya dilakukan resampling dengan menggunakan algoritma ADASYN. Algoritma ini bekerja dengan menghasilkan nilai baru dengan melihat distribusi kelas dengan bergantung kepada seberapa susah suatu kelas untuk dipelajari atau seberapa jauh perbandingan kelas tersebut dengan kelas lain(Satapathy et al., 2023).

Pada Gambar 2 dan 3 di bawah merupakan persebaran dari data yang ada di dalam data set terhadap kelasnya. Kelas 0 merupakan emosi marah, kelas 1 emosi takut, kelas 2 emosi senang, kelas 3 emosi sedih dan kelas 4 emosi netral. Pada Gambar 2 menjelaskan bahwa terjadinya perbedaan jumlah data yang signifikan di dalam data set di mana kelas netral memiliki jumlah data sebesar 423 data dan kelas lainnya hanya memiliki data sebesar 282 data.

Perbedaan data yang signifikan ini akan menyebabkan model kesusahan untuk memahami data baru karena model hanya menghafal pola. Oleh karena itu dilakukannya resampling ADASYN terhadap data latih seperti terlihat pada Gambar 3. Hasil resampling ini menjadikan data di kelas 0 menjadi sebanyak 405 data, kelas 1 menjadi 456 data, kelas 2 menjadi 467 data, kelas 3 menjadi 466 data dan kelas 4 tetap berada di 423 data.



Gambar 2 Sebelum Oversampling ADASYN



Gambar 3 Sesudah Oversampling ADASYN

2.6. Klasifikasi CNN

Metode CNN yang digunakan di dalam penelitian ini memiliki beberapa lapisan yang digunakan.

1. Masking Layer (1 Lapisan)
2. Lapisan Konvolusi (5 Lapisan)
3. Lapisan Aktivasi (8 Lapisan)
4. Lapisan Pooling (5 Lapisan)
5. Lapisan Dropout (4 Lapisan)
6. Flattening (1 Lapisan)
7. Lapisan Dense (3 Lapisan)

Pada proses pertama, data ekstraksi fitur Spectrogram akan masuk ke dalam lapisan masking. Pada dasarnya, lapisan masking ini digunakan untuk mengatasi perbedaan panjang data maupun untuk mengatasi nilai nol dari padding yang dilakukan (Schneider et al., 2024). Selanjutnya, proses akan masuk kedalam Lapisan konvolusi yang mana jika dilihat secara matematis dapat dihasilkan rumus berikut,

$$FM[i]_{j,k} = \sum_m \sum_n F_{[m,n]} * N_{[j-m,\ k-n]} + bF$$

dimana :

FM[i] : Matriks Feature Map ke-i
N : Matriks Masukan
F : Matriks Konvolusi
bF : Nilai Bias
j,k : Posisi nilai piksel dalam matriks masukan
m,n : Posisi nilai piksel pada matriks konvolusi

Lapisan konvolusi bekerja dengan menggunakan filter untuk menghitung nilai konvolusi dari gambar sehingga menghasilkan ekstraksi fitur yang lebih kecil. Filter ini akan berjalan dari kiri ke kanan dan dari atas ke bawah sesuai ukuran filter yang ditentukan terhadap gambar (Purwono et al., 2022) .

Kemudian hasil konvolusi tersebut akan masuk ke dalam lapisan aktivasi yang mana di dalam penelitian ini, aktivasi yang digunakan adalah *leaky_relu.* Lapisan aktivasi digunakan untuk mengenalkan non-linearitas kedalam metode yang dibuat sehingga metode tersebut mungkin untuk memahami pola (Nanni et al., 2020) . Selanjutnya, proses akan masuk ke dalam lapisan pooling yang mana di dalam lapisan pooling ini fitur akan digabungkan atau dikelompokkan menjadi ukuran kecil tergantung dari ukuran yang digunakan sehingga akan mengurangi ukuran dimensi dari fitur dan memberikan metode CNN yang dibuat untuk tahan dari pergeseran data (Nirthika et al., 2022) .

Setelahnya masuk ke dalam lapisan dropout Lapisan droput ini dilakukan untuk meregulasi metode CNN yang dibuat dengan mematikan neuron secara acak di dalam CNN sehingga mengurangi terjadinya *overfitting* (Mao & Liu, 2023) . Setelah dari lapisan dropout, akan dilakukan flattening. Flatenning dilakukan untuk mengubah data fitur menjadi satu dimensi vektor

yang mampu ditangkap oleh Dense layer (Chen et al., 2022) .Sebelum pada akhirnya masuk ke lapisan terakhir(dense) Lapisan dense bekerja dengan melakukan kalkulasi untuk menghasilkan prediksi. Secara sederhananya, lapisan ini mengambil seluruh data hasil flattening, memasukkannya ke dalam seluruh neuron sehingga mampu memberikan prediksi global (Wang et al., 2021), lapisan ini dapat dilihat secara matematis sebagai berikut,

$$hid_i = \sum_{j=1}^{n} X_j * V_{j,i} + V_0$$

dimana :

hid_i : Masukan untuk *hidden layer* ke – i
X_j : node X ke - j
$V_{j,i}$: Bobot V untuk X_j
V_0 : Nilai Bias V untuk hid_i

2.7. Data Latih

Data latih yang digunakan di dalam penelitian model ini di dapat dengan melakukan train test split dengan ukuran data latih sebesar 80% dan data uji sebesar 20 %. Untuk melihat persebaran datanya, dapat dilihat pada Gambar 2 Sebelum *Oversampling* ADASYN di mana di dalam gambar tersebut menjelaskan bahwa data latih memiliki jumlah data total sebesar 1551 data.

2.8. Randomize Search

Pada penelitian ini dilakukan randomized search. Randomized Search merupakan metode optimasi otomatis untuk mengatasi berbagai macam parameter–parameter yang akan digunakan di dalam model yang dibuat (Rimal et al., 2024). Hasil dari randomized search akan menjadi hasil terbaik dengan nilai parameter paling optimal berdasarkan arsitektur model.

2.9. Model Terbaik

Model terbaik didapatkan dari hasil optimasi parameter dalam tingkat pembelajaran, tingkat pemutusan, jenis kernel, bobot peluruhan, metode optimasi, jumlah epoch dan batch yang dilakukan pada tahap randomized search

2.10. Data Uji

Data uji yang digunakan dalam penelitian ini juga di dapat sebagai hasil dari dilakukannya train test split, di mana jumlah data uji adalah 517 data suara. Jika baik data uji maupun data latih di gabung maka nominal data adalah sebanyak 2068 data suara

2.11. Klasifikasi

Klasifikasi dilakukan setelah model telah belajar dari data latih dan mengaplikasikannya pengetahuan yang di dapat terhadap data uji

2.12. Confusion Matrix dan ROC Curve

Confusion Matrix merupakan metode yang digunakan untuk melakukan perhitungan akurasi. Metode ini bekerja dengan memasukkan hasil klasifikasi ke dalam empat jenis kelas. Benar Positif (TP), Benar Negative (TN), Salah Positif (FP), dan Salah Negatif (FN) (Riehl et al., 2023) . Nilai dari Benar Positif (TP) adalah representasi data positif yang terklasifikasi benar sementara itu, Benar Negative (TN) adalah representasi data negatif yang terklasifikasi benar (Heydarian et al., 2022). Begitu juga dengan Salah Positif (FP) menunjukkan data positif terklasifikasi sebagai salah, dan Salah Negatif (FN) menujukan data negatif terklasifikasi salah.

Sementara itu, kurva ROC merupakan metode yang digunakan untuk menunjukkan performa metode CNN dalam suatu ambang batas. Semakin banyak data yang berada di atas dari ambang batas maka semakin bagus suatu model begitu juga sebaliknya, semakin banyak data di bawah atau sama dengan ambang batas maka semakin jelek model tersebut (Zeng, 2020).

3. Hasil

Penelitian ini menggunakan *Randomized Search* dari parameter yang ada dengan keterbatasan 1551 data latih dan penggunaan augementasi ke seluruh data suara serta dengan kelas suara yang telah dilakukan resampling melalui ADASYN akibat tidak seimbangnya kelas suara. Tabel 1 Hasil Cross Validation memberikan penjelasan mengenai parameter apa yang digunakan di dalam pelatihan model seperti tingkat pembelajaran, tingkat pemutusan, jenis kernel, bobot peluruhan, metode optimasi, jumlah epoch dan batch. Sementara itu, Mean val menjelaskan nilai rata–rata akurasi validasi yang diperoleh suatu model dengan parameter yang ditentukan sebagai bentuk gambaran kinerja model dalam memahami data suara dan STD Val memberikan penjelasan mengenai tingkat deviasi dari model yang mana menunjukkan konsistensi kinerja model dimana semakin kecil nilai STD Val maka semakin konsisten model tersebut dalam memahami data suara.

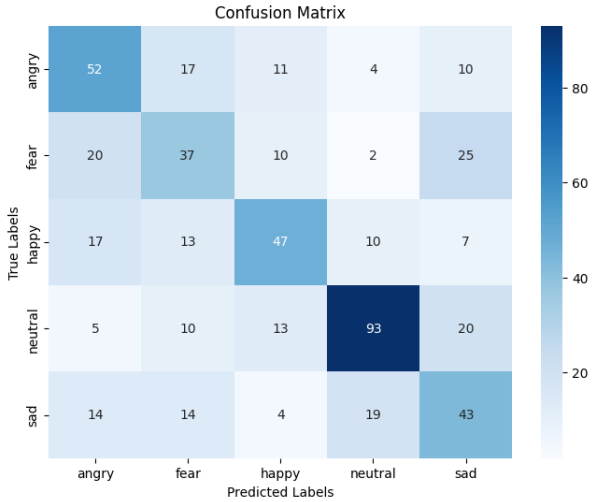
Tabel 1 Hasil Cross Validation

Parameter	Mean Val	STD Val
{'model__weight_decay': 1e-07, 'model__optimizer': 'adamw', 'model__learning_rate': 0.001, 'model__kernel_initializer': 'glorot_uniform', 'model__dropout_rate': 0.6, 'epochs': 120, 'batch_size': 16}	0.3074124602091 86	0.138601714526 1076
{'model__weight_decay': 1e-07, 'model__optimizer': 'adamw', 'model__learning_rate': 0.001, 'model__kernel_initializer': 'he_normal', 'model__dropout_rate': 0.5, 'epochs': 100, 'batch_size': 48}	0.7448840381991 815	0.080417876191 7074
{'model__weight_decay': 1e-08, 'model__optimizer': 'rmsprop', 'model__learning_rate': 0.001, 'model__kernel_initializer': 'glorot_uniform', 'model__dropout_rate': 0.6, 'epochs': 120, 'batch_size': 32}	0.7216916780354 707	0.077462862573 4154
{'model__weight_decay': 1e-07, 'model__optimizer': 'rmsprop', 'model__learning_rate': 0.001, 'model__kernel_initializer': 'glorot_uniform', 'model__dropout_rate': 0.5, 'epochs': 100, 'batch_size': 24}	0.7157798999545 247	0.065050226158 9673
{'model__weight_decay': 1e-08, 'model__optimizer': 'rmsprop', 'model__learning_rate': 0.001, 'model__kernel_initializer': 'glorot_uniform', 'model__dropout_rate': 0.5, 'epochs': 90, 'batch_size': 16}	0.6962255570713 961	0.067722964434 6781
{'model__weight_decay': 1e-08, 'model__optimizer': 'rmsprop', 'model__learning_rate': 0.01, 'model__kernel_initializer': 'he_normal', 'model__dropout_rate': 0.5, 'epochs': 100, 'batch_size': 24}	0.2037289677125 9663	0.016000556694 6422
{'model__weight_decay': 1e-08, 'model__optimizer': 'rmsprop', 'model__learning_rate': 0.0001, 'model__kernel_initializer': 'he_normal', 'model__dropout_rate': 0.5, 'epochs': 120, 'batch_size': 32}	0.4997726239199 6363	0.021493955976 5283
{'model__weight_decay': 1e-05, 'model__optimizer': 'rmsprop', 'model__learning_rate': 0.01, 'model__kernel_initializer': 'he_normal', 'model__dropout_rate': 0.5, 'epochs': 100, 'batch_size': 24}	0.2050932241928 1492	0.006333964655 3815
{'model__weight_decay': 1e-08, 'model__optimizer': 'rmsprop', 'model__learning_rate': 0.0001, 'model__kernel_initializer': 'he_normal', 'model__dropout_rate': 0.6, 'epochs': 90, 'batch_size': 32}	0.3601637107776 2627	0.015313741009 9888
{'model__weight_decay': 1e-05, 'model__optimizer': 'adamw', 'model__learning_rate': 0.0001, 'model__kernel_initializer': 'he_normal', 'model__dropout_rate': 0.6, 'epochs': 120, 'batch_size': 32}	0.3756252842201 0005	0.054528557660 295

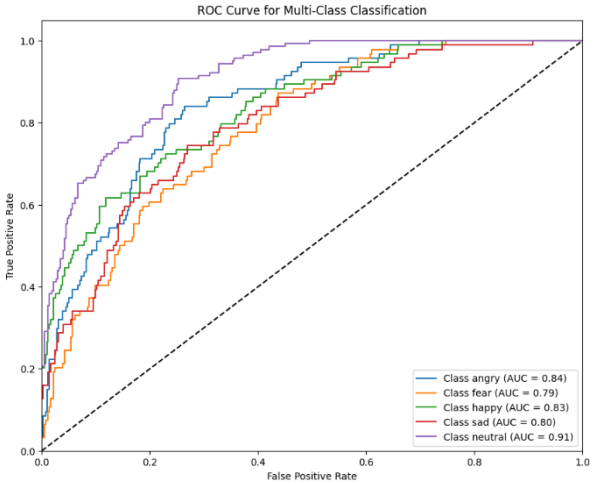
Berdasarkan Tabel 1 Hasil Cross Validation perolehan parameter model yang paling baik adalah dengan Best Parameters: {'model__weight_decay': 1e-07, 'model__optimizer': 'adamw', 'model__learning_rate': 0.001, 'model__kernel_initializer': 'he_normal', 'model__dropout_rate': 0.5, 'epochs': 100, 'batch_size': 48}, nilai mean memiliki Score sebesar 0.7448840381991815 dan nilai STD Val sebesar 0.0804178761917074. Model dengan parameter ini selanjutnya akan dilakukan pengujian kembali untuk menghasilkan nilai akurasi, Confusion Matrix (CF) dan ROC Curve.

Model ini menghasilkan nilai akurasi umum untuk data latih sebesar 75.85% dan untuk data uji sebesar 51.64%. Dengan rendahnya nilai akurasi pada data uji dan cukup tingginya akurasi data latih, model mengalami fenomena yang dinamakan sebagai overfitting di mana model hanya

menghafal pola yang ada dan kesusahan dalam memahami data baru atau data uji. Overfitting ini dapat terjadi akibat kurang kuatnya metode resampling yang digunakan atau kesalahan dalam augmentasi yang menyebabkan nilai di dalam fitur menjadi kurang stabil. Untuk lebih jelasnya, Gambar 4 *Confusion Matrix* menggambarkan persebaran salah dan benarnya prediksi dari model yang dihasilkan. Dari gambar yang ada kotak dengan warna biru pada posisi diagonal kiri menunjukkan prediksi data benar dengan kelasnya di mana kotak lain menunjukkan kesalahan prediksi. Kesalahan prediksi dengan nilai besar terjadi pada kelas takut di prediksi sebagai marah sebesar 20 data dan sedih sebesar 25 data. Hal yang sama terjadi pada kelas netral di prediksi sebagai kelas sedih sebesar 20 data.



Gambar 4 Confusion Matrix Best Model



Gambar 5 ROC Curve Best Model

Pada Gambar 5 *Roc Curve Best Mo*del, dihasilkan nilai prediksi per kelas yang merefleksi hasil CF. Kelas takut memiliki prediksi kelas sebesar 0.79 dilihat dalam CF terjadi dua kesalahan prediksi dengan jumlah salah yang besar yaitu data terprediksi sebagai kelas marah dan netral. Begitu juga sebaliknya, Kelas netral memiliki prediksi paling bagus sebesar 0.91 dan di lihat dalam CF kelas netral di prediksi benar sebesar 93 data. Dengan demikian, Model yang dibuat dengan parameter yang di ujikan di dalam Tabel 1 Hasil Cross Validation dan menggunakan metode CNN dengan fitur spectrogram memiliki kemampuan yang kurang baik untuk mendeteksi emosi suara bahkan ketika augmentasi sudah diterapkan keseluruh data untuk menyimulasikan kondisi di dunia nyata.

4. Kesimpulan

Penelitian ini menghasilkan metode CNN dengan tingkat akurasi umum untuk data latih sebesar 75.85% dan data uji sebesar 51.64%. serta dengan nilai kurva ROC yang menunjukkan kemampuan model dalam mengklasifikasikan data ke dalam kelasnya masing–masing sebesar 0.84 untuk kelas marah, 0.79 untuk kelas takut, 0.83 untuk kelas senang, 0.80 untuk kelas sedih, dan 0.91 untuk kelas netral. Dari data akurasi yang ada, model yang dihasilkan tidak mampu memprediksi emosi berdasarkan suara dengan baik akibat overfitting. Peneliti beranggapan overfitting yang terjadi di dalam penelitian ini diakibatkan kurang seimbangnya data suara bahkan ketika sudah dilakukan resampling atau kesalahan dalam melakukan augmentasi yang menyebabkan data suara menjadi sangat tidak cocok untuk dilakukan prediksi.

References

Aini, Y. K., Santoso, T. B., & Dutono, T. (2021). Pemodelan CNN Untuk Deteksi Emosi Berbasis Speech Bahasa Indonesia. *Jurnal Komputer Terapan*, *7*(1), 143–152. https://doi.org/10.35143/jkt.v7i1.4623

Alluhaidan, A. S., Saidani, O., Jahangir, R., Nauman, M. A., & Neffati, O. S. (2023). Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network. *Applied Sciences (Switzerland)*, *13*(8). https://doi.org/10.3390/app13084750

Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., Modi, K., & Ghayvat, H. (2021). Cnn variants for computer vision:

History, architecture, application, challenges and future scope. In *Electronics (Switzerland)* (Vol. 10, Issue 20). https://doi.org/10.3390/electronics10202470

Chen, P. Y., Zhang, X. H., Wu, J. X., Pai, C. C., Hsu, J. C., Lin, C. H., & Pai, N. S. (2022). Automatic Breast Tumor Screening of Mammographic Images with Optimal Convolutional Neural Network. *Applied Sciences (Switzerland)*, *12*(8). https://doi.org/10.3390/app12084079

Deepa, B., & Ramesh, K. (2022). Epileptic seizure detection using deep learning through min max scaler normalization. *International Journal of Health Sciences*. https://doi.org/10.53730/ijhs.v6ns1.7801

Gencyilmaz, I. Z., & Karaoğlan, K. M. (2024). Optimizing Speech to Text Conversion in Turkish: An Analysis of Machine Learning Approaches. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, *13*(2), 492–504. https://doi.org/10.17798/BITLISFEN.1434925

George, S. M., & Muhamed Ilyas, P. (2024). A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. *Neurocomputing*, *568*, 127015. https://doi.org/10.1016/J.NEUCOM.2023.127015

Heydarian, M., Doyle, T. E., & Samavi, R. (2022). MLCM: Multi-Label Confusion Matrix. *IEEE Access*, *10*. https://doi.org/10.1109/ACCESS.2022.3151048

Hu, S., Chu, Y., Wen, Z., Zhou, G., Sun, Y., & Chen, A. (2023). Deep learning bird song recognition based on MFF-ScSEnet. *Ecological Indicators*, *154*. https://doi.org/10.1016/j.ecolind.2023.110844

Juslin, P., & Scherer, K. (2008). Speech emotion analysis. *Scholarpedia*, *3*(10). https://doi.org/10.4249/scholarpedia.4240

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, *53*(8). https://doi.org/10.1007/s10462-020-09825-6

Li, J., Zhang, X., Huang, L., Li, F., Duan, S., & Sun, Y. (2022). Speech Emotion Recognition Using a Dual-Channel Complementary Spectrogram and the CNN-SSAE Neutral Network. *Applied Sciences (Switzerland)*, *12*(19). https://doi.org/10.3390/app12199518

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

Mao, Y., & Liu, Y. (2023). Pet dog facial expression recognition based on convolutional neural network and improved whale optimization algorithm. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-30442-0

Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. In *Array* (Vol. 16). https://doi.org/10.1016/j.array.2022.100258

Nanni, L., Lumini, A., Ghidoni, S., & Maguolo, G. (2020). Stochastic selection of activation layers for convolutional neural networks. *Sensors (Switzerland)*, *20*(6). https://doi.org/10.3390/s20061626

Nirthika, R., Manivannan, S., Ramanan, A., & Wang, R. (2022). Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study. In *Neural Computing and Applications* (Vol. 34, Issue 7). https://doi.org/10.1007/s00521-022-06953-8

Ozcelik, S. T. A., Uyanık, H., Deniz, E., & Sengur, A. (2023). Automated Hypertension Detection Using ConvMixer and Spectrogram Techniques with Ballistocardiograph Signals. *Diagnostics*, *13*(2). https://doi.org/10.3390/diagnostics13020182

Purwono, Ma'arif, A., Rahmaniar, W., Fathurrahman, H. I. K., Frisky, A. Z. K., & Haq, Q. M. U. (2022). Understanding of Convolutional Neural Network (CNN): A Review. *International Journal of Robotics and Control Systems*, *2*(4). https://doi.org/10.31763/ijrcs.v2i4.888

Rahmadani, S., Rahayu, C. S., Salim, A., & Cahyo, K. N. (2022). DETEKSI EMOSI BERDASARKAN WICARA MENGGUNAKAN DEEP LEARNING MODEL. *Jurnal Informatika Teknologi Dan Sains (Jinteks)*, *4*(3), 220–224.

https://doi.org/10.51401/JINTEKS.V4I3.1952

Riehl, K., Neunteufel, M., & Hemberg, M. (2023). Hierarchical confusion matrix for classification performance evaluation. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, *72*(5). https://doi.org/10.1093/jrsssc/qlad057

Rimal, Y., Sharma, N., & Alsadoon, A. (2024). The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms. *Multimedia Tools and Applications*, *83*(30). https://doi.org/10.1007/s11042-024-18426-2

Salehi, A. W., Khan, S., Gupta, G., Alabduallah, B. I., Almjally, A., Alsolai, H., Siddiqui, T., & Mellit, A. (2023). A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope. In *Sustainability (Switzerland)* (Vol. 15, Issue 7). https://doi.org/10.3390/su15075930

Satapathy, S. K., Mishra, S., Mallick, P. K., & Chae, G. S. (2023). ADASYN and ABC-optimized RBF convergence network for classification of electroencephalograph signal. *Personal and Ubiquitous Computing*, *27*(3). https://doi.org/10.1007/s00779-021-01533-4

Schneider, M., Greifzu, N., Wang, L., Walther, C., Wenzel, A., & Li, P. (2024). An end-to-end machine learning approach with explanation for time series with varying lengths. *Neural Computing and Applications*, *36*(13). https://doi.org/10.1007/s00521-024-09473-9

Tanudjaja, F. J., Puspaningrum, E. Y., & Via, Y. V. (2023). Klasifikasi Jenis Emosi Melalui Ucapan Menggunakan Metode Convolutional Neural Network : Klasifikasi Jenis Emosi Melalui Ucapan. *Teknologi: Jurnal Ilmiah Sistem Informasi*, *13*(2), 1–11. https://doi.org/10.26594/TEKNOLOGI.V13I2.3740

Wang, Z., Liu, Q., Chen, H., & Chu, X. (2021). A deformable CNN-DLSTM based transfer learning method for fault diagnosis of rolling bearing under multiple working conditions. *International Journal of Production Research*, *59*(16). https://doi.org/10.1080/00207543.2020.1808261

Zeng, G. (2020). On the confusion matrix in credit scoring and its analytical properties. *Communications in Statistics - Theory and Methods*, *49*(9). https://doi.org/10.1080/03610926.2019.1568485