E-ISSN 2808-5841
P-ISSN 2808-7283

Speech Emotion Recognition (SER) dengan Metode Bidirectional LSTM

Nicholas Juan Kalvin Pradiptamurty [1], Hafiyyah Khayyiroh Shafro[2], Mohammad Sihabudin Al Qurtubi[3], Giovanny Alberta Tambahjong[4], Qothrotunnidha' Almaulidiyah[5], Maryamah Maryamah[6]
[1, 2, 3, 4, 5] *Teknologi Sains Data, Universitas Airlangga*
[1] nicholas.juan.kalvin-2020@ftmm.unair.ac.id
[2] hafiyyah.khayyiroh.afro-2020@ftmm.unair.ac.id
[3] mohammad.sihabudin.al-2020@ftmm.unair.ac.id
[4] giovanny.alberta.tambahjong-2020@ftmm.unair.ac.id
[5] qothrotunnidha.almaulidiyah-2020@ftmm.unair.ac.id
[6] maryamah@ftmm.unair.ac.id
Corresponding author email: maryamah@ftmm.unair.ac.id

Abstract: *Emotions are a part of humans as a form of response to experienced events. Emotion analysis or known as speech emotion recognition (SER) is a field many researchers are interested in because voice recognition systems can assist in criminal investigations, monitoring, and detection of potentially dangerous events, and assisting the health care system. Therefore, this study proposes the detection of SER using the Bidirectional Long short-term memory (Bi-LSTM) model approach. The dataset used was scraped on the YouTube platform. The dataset is manually labeled then feature extraction is performed using the Mel Frequency Cepstral Coefficients (MFCC). The experiment using the Bi-LSTM method has an AUC ROC value of 0.97 and an f1-score value of 0.878. Based on these results, it can be concluded that the performance of the proposed method succeeded in predicting SER better than other comparison methods. This model also proved to be more precise in classifying human voices based on four types of emotions, namely happy, sad, angry, and neutral.*
Keywords: *Speech Emotion Recognition, Bidirectional Long short-term memory (Bi-LSTM), Audio Classification*
Abstrak: Emosi merupakan bagian tak terpisahkan dari setiap manusia sebagai bentuk respon terhadap suatu peristiwa yang dialami. Analisis emosi atau dikenal dengan *speech emotion recognition* (SER) menjadi bidang penelitian yang diminati oleh banyak peneliti karena sistem pengenalan suara dapat membantu dalam investigasi kriminal, pengawasan dan deteksi peristiwa yang berpotensi berbahaya, serta membantu sistem perawatan kesehatan. Oleh karena itu, penelitian ini mengusulkan deteksi SER dengan pendekatan model Bidirectional Long short-term memory (Bi-LSTM). Dataset yang digunakan diambil dengan metode *scraping* pada platform youtube. Selanjutnya dataset dilakukan manual label dan dilakukan ekstraksi fitur menggunakan *Mel Frequency Cepstral Coefficients* (MFCC). Metode Bi-LSTM menghasilkan nilai ROC AUC sebesar 0,97 dan nilai *F1-score* sebesar 0,878. Berdasarkan hasil tersebut maka dapat disimpulkan bahwa kinerja metode yang diusulkan, yakni model Bi-LSTM berhasil memprediksi SER lebih baik dibandingkan dengan metode lainnya. Model ini pun terbukti lebih tepat dalam mengelompokkan suara manusia berdasarkan empat jenis emosi yakni senang, sedih, marah, dan netral.
Kata kunci: Speech Emotion Recognition, Bidirectional Long Short-Term Memory (Bi-LSTM), Klasifikasi Audio

I. PENDAHULUAN

Emosi adalah bagian yang tidak terhindarkan dari setiap pribadi seseorang pada saat berkomunikasi [1]. Bentuknya dapat diekspresikan dengan berbagai macam cara yang terkadang tidak dapat diketahui secara langsung [2]. Pada dasarnya setiap manusia mempunyai emosi sebagai bentuk implementasi atau respon yang melibatkan perilaku dan fisiologis yang digunakan untuk menangani sebuah peristiwa yang dialami. Emosi dapat dikatakan sebagai faktor yang mudah diketahui oleh sesama manusia dalam berkomunikasi dan sebagai sarana menambah cita rasa hidup dengan memperkenalkan cara mengekspresikan perasaan mereka dalam bentuk komunikasi [3].

Analisis emosi manusia telah menjadi topik penelitian dalam berbagai disiplin ilmu, seperti Ilmu Kognitif, Psikologi. Berkat difusi media sosial ini dapat menarik minat para ilmuwan

komputer atau mesin cerdas [4]. Akibat perkembangan zaman, interaksi kerjasama manusia-mesin semakin sering ditemui sehingga kebutuhan akan sistem dalam mengerti manusia semakin meningkat, salah satu contohnya adalah mesin dipaksa untuk dapat memahami keadaan emosi dari suara manusia. Ketika mesin dapat mendeteksi emosi manusia dari interaksi yang berlangsung, hal itu dapat berguna untuk mengetahui kondisi psikologis atau emosi dari manusia tersebut [5]. Selain itu, pengenalan emosi yang dilakukan mesin sangat berguna untuk aplikasi yang membutuhkan interaksi manusia-mesin seperti *customer service* [6]. *Customer service* dalam pelayanannya memerlukan respon yang disesuaikan berdasarkan emosi yang terdeteksi. Oleh karena itu, dikenalkan bidang penelitian yang baru, yaitu *speech emotion recognition* (SER).

Speech emotion recognition adalah sebuah sistem yang dapat mengklasifikasikan perasaan atau emosi seseorang dari gaya berbicaranya. Berdasarkan teori *speech emotion analysis*, emosi seseorang dapat dirasakan secara nonverbal dari perubahan ritme respirasi, tensi otot yang menggetarkan suara dan mengganti karakteristik akustik, dan lain-lain [7]. *Speech emotion recognition* diyakini dapat meningkatkan kinerja sistem pengenalan suara sehingga membantu dalam investigasi kriminal, pengawasan dan deteksi peristiwa yang berpotensi berbahaya, serta membantu sistem perawatan kesehatan [8]. Untuk dapat mengenali emosi secara efektif dan akurat dari audio ucapan seseorang, audio tersebut perlu diekstraksi dari audio analog ke audio digital sehingga dapat diproses lebih lanjut. Akan tetapi, diferensiasi dari emosi terbukti susah untuk ditentukan karena sifat natural suara yang kompleks. Beberapa penelitian terkait SER telah dilakukan dengan menggunakan metode *Hidden Markov Model* (HMM) dan MFCC berhasil menentukan 4 dari 7 jenis emosi yang digunakan dengan akurasi lebih dari 50% [5], penelitian lain menerapkan algoritma *boosting* untuk SER dalam bahasa Indonesia dan memperoleh akurasi sebesar 65% [9], dan pemodelan dengan berbagai metode RNN menghasilkan Bi-LSTM sebagai model terbaik dalam menghasilkan error yang paling kecil di antara metode RNN lainnya [10].

Salah satu metode ekstraksi suara yang banyak digunakan dalam bidang *speech technology* yakni rekognisi suara maupun ucapan adalah MFCC. MFCC sendiri digunakan untuk melakukan ekstraksi fitur dengan mendapatkan nilai-nilai tertentu sebagai konsep konversi sinyal suara menjadi beberapa parameter tertentu. Penerapan ekstraksi fitur MFCC berdasarkan penelitian terdahulu yang menerapkan ekstraksi fitur MFCC terbukti sangat baik dengan taraf keakuratan tinggi dalam mengekstraksi fitur suara yang merepresentasikan suara manusia, terutama pada *speech recognition* [11].

Berdasarkan permasalahan di atas, paper ini mengusukan *speech emotion recognition* (SER) menggunakan metode Bi-LSTM dengan ekstraksi fitur menggunakan MFCC. Selain menggunakan metode Bi-LSTM, akan digunakan juga metode LSTM dan CNN untuk digunakan perbandingan dalam menghasilkan metode mana yang paling baik. Dalam mengatasi masalah terkait *imbalanced*, akan digunakan *class weights* sebagai metode dalam mengatasi kelas yang *imbalanced* agar tidak terlalu mengubah informasi data audio yang diperoleh. Oleh karena itu, penelitian ini bertujuan mendapatkan informasi dari *audio mining* untuk mendeteksi SER dengan tepat dan akurat menggunakan pendekatan model Bi-LSTM, LSTM, dan CNN kemudian dievaluasi menggunakan satuan ukur nilai akurasi.

II.

II. METODE PENELITIAN

2.1. Sumber Data

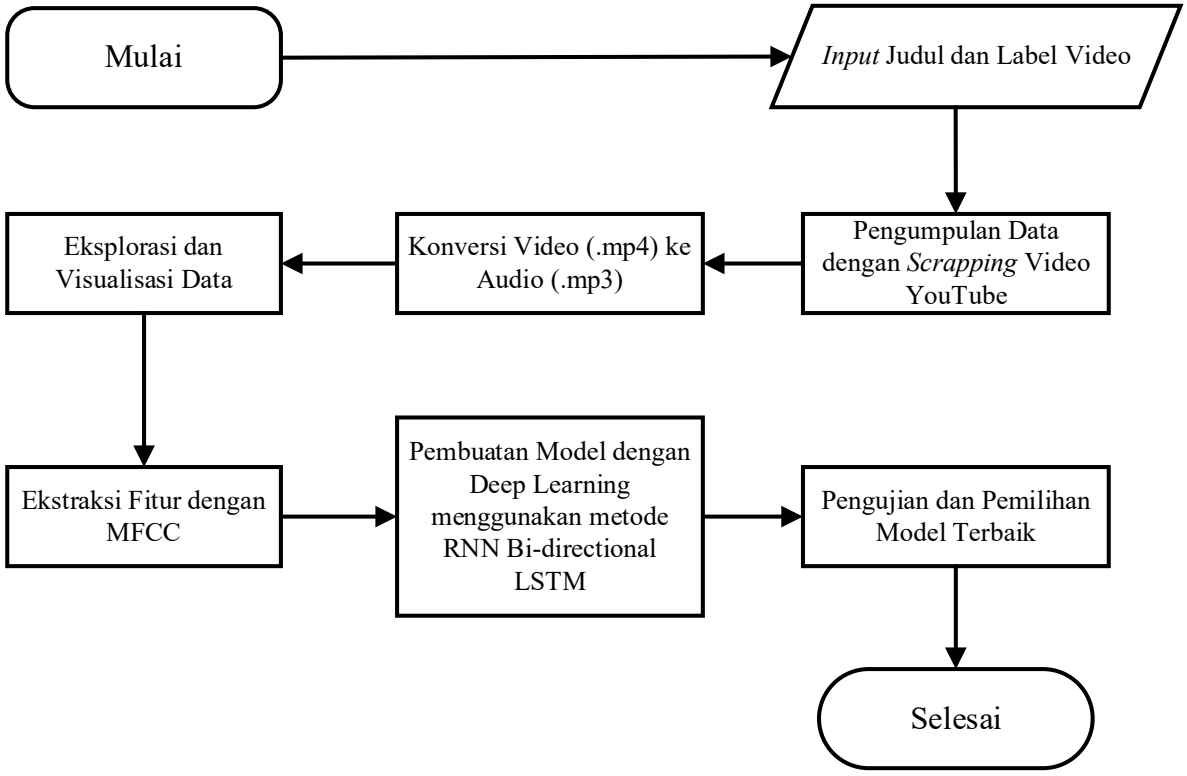
Data *audio* didapatkan dengan melakukan *scraping* atau mengambil data yang tersedia secara *online*. Sumber data yang digunakan adalah dari platform media sosial *YouTube* dimana tersedia konten video dengan jumlah yang sangat banyak. Platform ini memberikan peneliti banyak pilihan data yang dapat digunakan. Peneliti harus jeli dalam memilih video dimana suara terdengar bersih dengan emosi yang jelas. Pemilihan video yang ambigu (sarkas, banyak orang, atau banyak suara lain) tidaklah

dipreferensikan. Ditemukan bahwa video terbaik yang digunakan untuk analisis adalah video bergenre esai video, komentar, atau video log (*vlog).*

Teknik pengambilan video adalah dengan membuat skrip otomasi menggunakan bahasa pemrograman *Python*, yang dilengkapi dengan library utama *yt-dlp, ffmpeg* dan *youtube-search.* Library lain yang digunakan untuk pemrosesan data pun juga digunakan. Cara kerja program *scraping* ini adalah dengan mencari video dari platform *YouTube* sesuai dengan judul yang ditentukan oleh peneliti yang dimasukkan ke dalam program. Setelah video ditemukan, program akan meminta peneliti untuk memilih label yang tepat yang merepresentasikan video tersebut. Setelah label ditentukan, barulah program akan mengunduh dan mensegmentasi video menjadi beberapa bagian audio masing-masing berdurasi 3 detik dan diletakan di folder sesuai dengan label yang telah ditentukan.

2.2. Metodologi Penelitian

Tahapan analisis dalam penelitian ini adalah studi literatur, penentuan metode dan pengumpulan data, eksplorasi dan visualisasi data, ekstraksi fitur, pembuatan model, dan evaluasi model. *Flowchart* tahapan penelitian dapat dilihat pada Gambar 1.



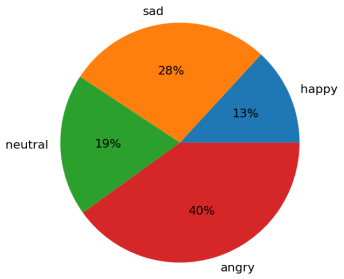
Gambar 1. *Flowchart* Penelitian

Adapun penjelasan tahapan analisis secara rinci adalah sebagai berikut:

1. Studi literatur dilakukan untuk menentukan pendekatan terbaik untuk mendapatkan sistem yang efektif dan efisien berdasarkan penelitian dan pendekatan yang telah dilakukan sebelumnya oleh peneliti lain. Peneliti mencari penelitian-penelitian sebelumnya terkait audio *mining* di GoogleScholar dengan kata kunci "Audio Mining". Setelah itu, menentukan topik untuk audio *mining* berdasarkan kemampuan dan ke-*familiar*-an topik yang dipilih;
2. Menentukan metode dan objek pengumpulan data yang dalam hal ini adalah *scrapping* video dari laman YouTube. Peneliti mencari dan mengunduh video yang relevan dengan pengekspresian emosi manusia ketika, marah, senang, netral, dan sedih. Video lalu

disegmentasi per 3 detik lalu diekspor menjadi file suara .mp3. Dalam tahap *scrapping*, peneliti membuat modul dalam format .py. Modul tersebut berfungsi untuk melakukan *scrapping* serta mengubahnya menjadi potongan-potongan audio berdurasi 3 detik yang terkumpul dalam folder khusus. Hal ini dimaksudkan agar dalam proses *scrapping* video, peneliti tidak perlu mengeksekusi kode rumit berulang kali dan hanya perlu meng-*import* modul yang telah dibuat lalu menginisiasinya. Pemotongan audio menjadi 3 detik dilakukan dengan tujuan agar dalam pemrosesan data audio tidak terlalu berat dan ketika audio dibagi menjadi beberapa segmen, tiap segmen memuat informasi yang berbeda-beda sehingga sistem mendeteksi data yang heterogen dari satu audio yang berdurasi lebih dari 3 detik. Dengan demikian, dapat dilakukan proses *train test split* dengan banyak data;

3. Eksplorasi dan visualisasi data (EVD) dilakukan untuk menentukan pendekatan yang lebih spesifik terhadap data yang diperoleh di tahap sebelumnya. Hal ini diharapkan akan memberikan ide akan parameter yang akan digunakan pada tahap ekstraksi fitur serta mengetahui informasi dari data yang diperoleh sehingga dapat melakukan analisis/pemodelan yang tepat. EVD yang dilakukan dalam penelitian ini adalah membuat plot *waveform*, plot *spectrum,* mencari nilai dominan dari *fast fourier transform* (FFT), dan plot *spectrogram* pada tiap kelas yang ada. *Pie chart* pada Gambar 2 dibuat untuk menunjukkan jumlah data dari masing-masing label.



Gambar 2. *Pie Chart* Label Emosi

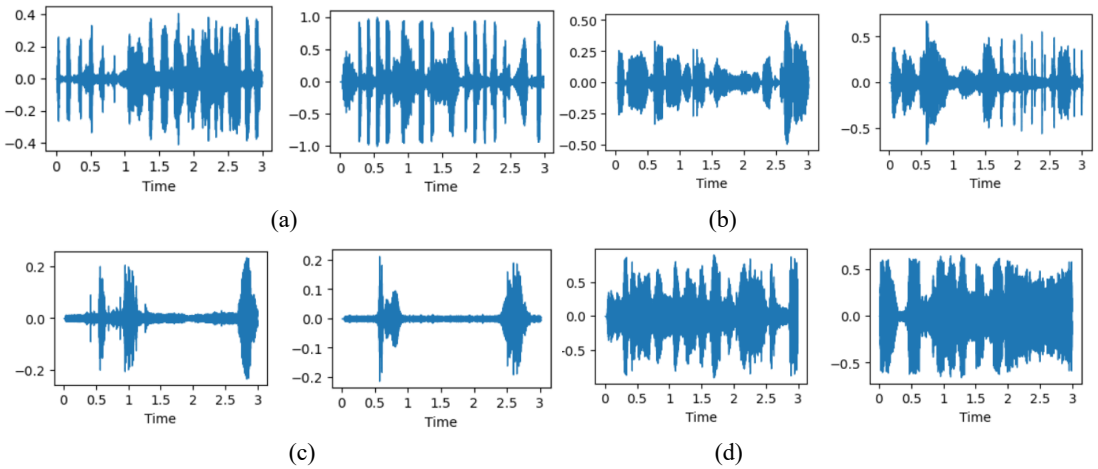
Dapat dilihat bahwa label angry merupakan label yang mempunyai jumlah data paling banyak di antara label yang lainnya, yakni sebesar 40%. Sedangkan untuk label happy merupakan label yang mempunyai jumlah data paling sedikit, yakni hanya sebesar 13% saja. Disebabkan distribusi jumlah data yang tidak merata maka dapat dikatakan bahwa terjadi *imbalanced* data dalam penelitian ini. Adapun rincian jumlah data dapat dilihat pada Tabel 1.

Tabel 1. Jumlah Data pada Tiap Kelas Emosi

Happy	Sad	Neutral	Angry	Jumlah Data
427	897	624	1.302	3.250

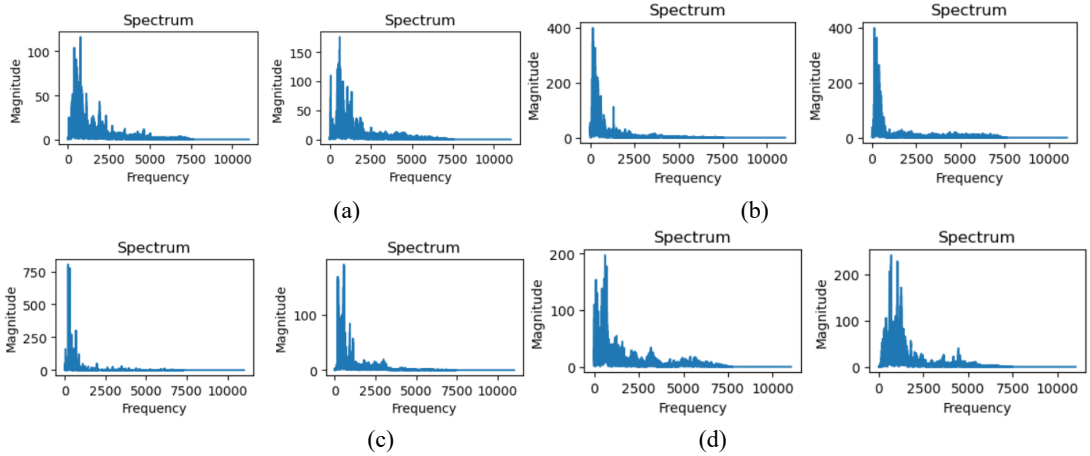
Selanjutnya dibuat *waveform* pada Gambar 3 untuk setiap emosi sehingga dapat memberikan informasi terkait amplitudo dari sinyal audio. Hasilnya adalah pola yang dibentuk oleh keempat *waveform* cukup berbeda. *Waveform happy* mempunyai pola dengan amplitudo yang seringkali naik dalam beberapa waktu. *Waveform neutral* mempunyai pola yang cenderung stabil. *Waveform sad* mempunyai pola yang cukup stabil, yakni kenaikan amplitudo hanya di awal dan akhir, dan amplitudo yang dibentuk cenderung mendekati garis 0. Adapun pola dengan amplitudo tinggi dicurigai akibat adanya *noise* tangisan pada sampel suara *sad*. *Waveform angry* mempunyai pola dengan amplitudo yang lebih sering naik dan kenaikan tersebut cukup konstan jika dibandingkan dengan *waveform happy*.

E-ISSN 2808-5841
P-ISSN 2808-7283



Gambar 3. *Waveform* (a) *happy*; (b) *neutral*; (c) *sad*; (d) *angry*

Audio *angry* mempunyai pola amplitudo tinggi yang lebih konstan di setiap waktunya jika dibandingkan dengan *waveform* emosi lainnya disebabkan ketika emosi yang dirasakan adalah marah, manusia cenderung mengeluarkan suara yang lebih keras dibandingkan suara biasanya. Begitu pun dengan audio *happy* yang jika dilihat amplitudo yang terlihat cenderung lebih banyak yang naik karena ketika emosi yang dirasakan adalah senang, manusia cenderung sedikit menaikkan suara bicara atau tertawa. Ketika manusia merasakan emosi neutral, suara yang diberikan biasa saja dalam artian tidak terlalu kecil dan tidak terlalu besar, sedangkan ketika manusia merasakan emosi *sad*, suara yang diberikan cenderung memberikan amplitudo yang lebih kecil dibandingkan yang lainnya disebabkan rata-rata manusia ketika merasakan sedih cenderung menangis atau diam. Selanjutnya dibuat *spectrum* pada Gambar 4 sebagai representasi visual dari sinyal audio dalam domain frekuensi sehingga dapat mengidentifikasi frekuensi dari sinyal audio berdasarkan empat jenis emosi manusia. Hasilnya adalah *spectrum happy* dan *angry* mempunyai magnitude dominan yang cenderung lebih banyak pada rentang frekuensi 0 – 2500 jika dibandingkan dengan *spectrum neutral* dan *sad*.



Gambar 4. *Spectrum* (a) *happy*; (b) *neutral*; (c) *sad*; (d) *angry*

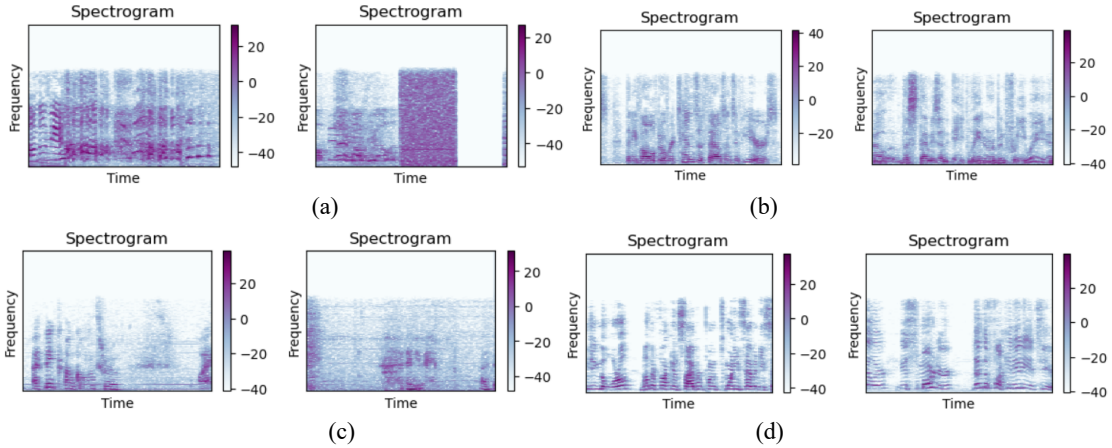
Untuk dapat menginterpretasikannya dengan jelas, peneliti mencari nilai rata-rata dominan frekuensi dari keseluruhan data tiap kelas. Hasil dari rata-rata tersebut akan dibagi dengan tiga yang merupakan durasi tiap audio yang kemudian akan menghasilkan rata-rata nilai Hz dari tiap kelas. Adapun nilai rata-rata dominan frekuensi dan rata-rata nilai Hz tiap kelas dapat dilihat pada Tabel 2.

Tabel 2. Nilai Rata-Rata Dominan Frekuensi dalam Hz

Happy	Sad	Neutral	Angry
19305,95	17237,37	16548,24	17640,45

Berdasarkan Tabel 2, label *happy* menghasilkan nilai 19305,95 Hz, *sad* menghasilkan nilai 17237,37 Hz, *neutral* menghasilkan nilai 16548,24 Hz, dan *angry* menghasilkan nilai 17640 Hz. Hal ini berarti dominan frekuensi yang paling besar dihasilkan oleh label *happy*, disusul dengan label *angry*, *sad*, dan *neutral*. Dengan demikian, dapat disimpulkan bahwa rata-rata dominan frekuensi yang dihasilkan label *happy* ternyata lebih besar dibandingkan rata-rata dominan frekuensi kelas yang lain.

Lalu, pembuatan *spectrogram* pada Gambar 5 dilakukan untuk menampilkan informasi spektrum frekuensi sinyal dalam bentuk warna sehingga lebih mudah dalam mengidentifikasi frekuensi dan waktu dari sinyal audio serta lebih mudah dalam menemukan pola dalam sinyal audio. Hasilnya adalah dapat dilihat bahwa *spectrogram happy* terlihat ada jarak antar *sample*, tetapi ada juga *spectrogram* yang mempunyai suara paling keras di tengah-tengah daripada suara sebelum dan sesudahnya. Pada label *sad* terdapat jarak yang seragam per *sample* diantara puncak-puncak frekuensi. Jarak tersebut dapat dijelaskan oleh pembicara yang mengambil waktunya untuk bernapas atau menghirup udara dari hidung. Frekuensi yang rata-rata rendah juga menjadi pola dari emosi sedih. Pada label *neutral* distribusi suara yang terdengar terlihat cukup merata dibandingkan dengan label lainnya. Hal ini berarti dalam kelas *neutral*, manusia cenderung berbicara tanpa henti dibandingkan ketika merasa senang, sedih, dan marah. *Spectogram* label *angry* juga terdapat jarak, diasumsikan jarak tersebut adalah ketika manusia menggunakan waktu tersebut untuk bernapas yang mungkin dilakukan untuk mengontrol emosinya;



Gambar 5. *Spectogram* (a) *happy*; (b) *neutral*; (c) *sad*; (d) *angry*

4. Ekstraksi fitur adalah tahap terpenting yang dilakukan pada penelitian ini disebabkan hasil akhir model akan ditentukan oleh seberapa baik data yang masuk ke dalam sistem. Ada banyak parameter yang diubah untuk mendapatkan fitur terbaik di antara kelas. Parameter tersebut ditentukan berdasarkan hasil dari tahap sebelumnya. Peneliti melakukan *label encoding* terhadap variabel-variabel yang digunakan. Hasil dari *label encoding* tersebut disimpan dalam format .npy. Lalu, peneliti melakukan ekstraksi fitur menggunakan metode MFCC dari data yang telah dilakukan standarisasi dan *balancing*.
5. Pembuatan model dengan arsitektur *deep learning* dilakukan setelah diperoleh data hasil ekstraksi fitur pada tahap sebelumnya. Model dibuat dengan metode RNN Bi-directional LSTM berdasarkan penelitian-penelitian sebelumnya yang menghasilkan model yang mendekati sempurna dengan menggunakan metode tersebut. Bidirectional LSTM merupakan tipe arsitektur *Recurrent Neural Network* (RNN) yang terdiri dari dua lapisan LSTM yang berbeda, yaitu lapisan LSTM maju (*forward* LSTM *layer*) dan lapisan LSTM mundur (*backward* LSTM *layer*). Arsitektur ini memungkinkan data masukan diolah dalam arah maju dan mundur, sehingga jaringan dapat menangkap karakteristik dan pola data yang sebelumnya diabaikan

oleh jaringan LSTM tradisional, pengolahan arah mundur pada jaringan Bi-LSTM memiliki keunggulan tersendiri, karena memungkinkan jaringan untuk belajar dari konteks data masa depan dan masa lalu. Dengan menangkap informasi dari kedua arah, jaringan dapat memodelkan ketergantungan dalam data masukan dengan efektif dan meningkatkan akurasi [12];

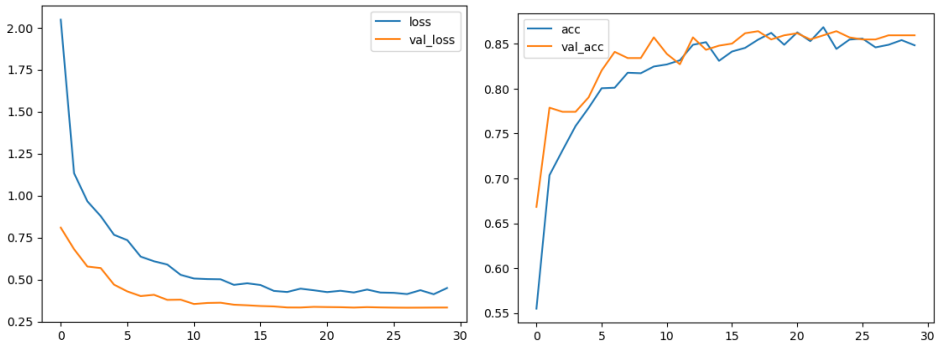
6. Tahap akhir adalah melakukan evaluasi berdasarkan akurasinya dan dilakukan prediksi langsung terhadap model yang telah dibuat. Prediksi dilakukan dengan merekam suara secara langsung dan disimpan dalam bentuk .mp3 yang setelahnya akan dimasukkan ke dalam model untuk dilakukan prediksi kelas emosi terhadap suara tersebut.

III. HASIL DAN PEMBAHASAN

Pada paper ini, jaringan Bi-LSTM digunakan untuk mempelajari fitur emosi dari klip audio mentah yang sudah diekstraksi fitur dengan MFCC serta membagi data menjadi dua set secara acak, yaitu set pelatihan dengan 80% data dan set pengujian dengan 20% data. Meskipun jaringan dalam *deep learning* dianggap sebagai "black box" *approach*, yakni cara kerjanya tidak begitu jelas, jaringan *deep learning* ini biasanya digunakan untuk menemukan algoritma yang dapat melakukan prediksi. Jaringan Bi-LSTM yang dirancang juga digunakan untuk kekuatan prediktifnya daripada kekuatan penjelas yang lemah. Untuk mengurangi kemungkinan atau jumlah *overfitting* pada penelitian, beberapa teknik diperkenalkan. *Overfitting* adalah salah satu alasan prediksi buruk untuk data sampel yang tidak terlatih. Ketika *overfitting* terjadi, model yang terlalu dipaksakan hanya mengingat data pelatihan daripada belajar untuk memprediksi dengan lebih baik. Ada banyak alasan terjadinya *overfitting*. Jika jaringan *deep learning* sangat kompleks, maka *overfitting* akan terjadi. Jika jaringan *deep learning overtrained*, *overfitting* juga akan timbul. Ketika derajat kebebasan model yang diadopsi dalam pelatihan jaringan terlalu banyak, maka kondisi *overfitting* akan ada. Oleh karena itu, dilakukan teknik normalisasi *batch, early stopping*, dan pemilihan model untuk mengatasi *overfitting*.

Dalam paper ini, hanya model-model terbaik yang dipilih untuk dicatat sebagai hasil yang tepat dan dapat diprediksi. Ketika akurasi validasi tidak lagi meningkat selama pelatihan model, itu menunjukkan bahwa model memiliki performa prediksi yang lebih superior (lihat Gambar 6). Dari gambar tersebut, dapat dilihat bahwa ketika akurasi validasi mencapai maksimum, akurasi pelatihan tidak mencapai maksimum. Ketika akurasi validasi menurun sementara akurasi pelatihan tetap meningkat, itu menunjukkan situasi *overfitting*. Oleh karena itu, pelatihan akan dihentikan dengan metode *early stopping*.

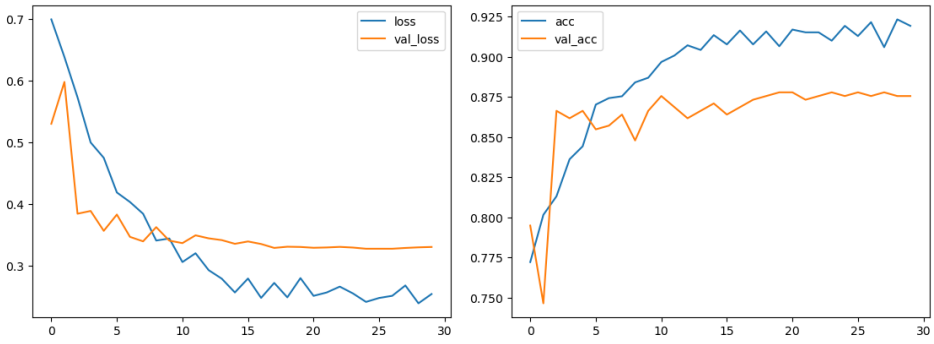
Pada grafik *loss* dan *accuracy* Gambar 6 dapat dilihat bahwa nilai loss dan *validation loss* dari model CNN berada di kisaran 0,5, sedangkan untuk nilai *accuracy* dan *validation accuracy* di kisaran 0,85. Disebabkan tidak adanya *gap* yang jauh antara garis kuning dan garis biru maka disimpulkan bahwa model yang dihasilkan tidak terjadi *overfit*.



(a) (b)

Gambar 6. Grafik (a) *loss*; (b) *accuracy* Model CNN

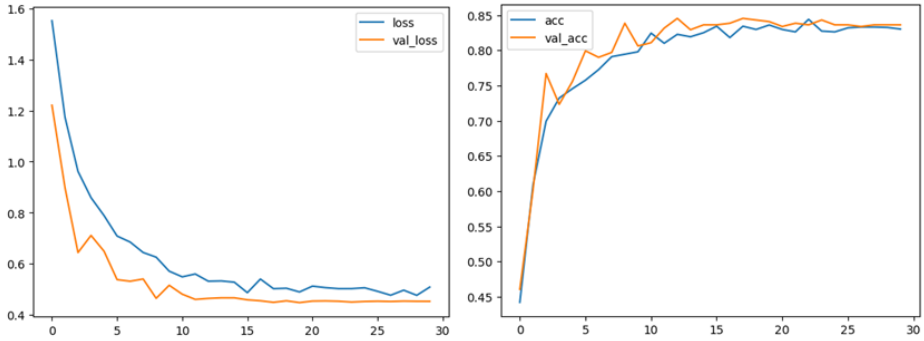
Pada grafik *loss* dan *accuracy* Gambar 7 dapat dilihat bahwa nilai *loss* dan *validation loss* dari model Bi-LSTM berada di kisaran 0,3, sedangkan untuk nilai *accuracy* dan *validation accuracy* dikisaran 0,88. Disebabkan tidak adanya *gap* yang jauh antara garis kuning dan garis biru maka disimpulkan bahwa model yang dihasilkan tidak terjadi *overfit*.



(a) (b)

Gambar 7. Grafik (a) *loss*; (b) *accuracy* Model Bi-LSTM

Pada grafik *loss* dan *accuracy* Gambar 8 dapat dilihat bahwa nilai *loss* dan *validation loss* dari model LSTM berada di kisaran 0,5, sedangkan untuk nilai *accuracy* dan *validation accuracy* di kisaran 0,83. Disebabkan tidak adanya *gap* yang jauh antara garis kuning dan garis biru maka disimpulkan bahwa model yang dihasilkan tidak terjadi *overfit*.



(a) (b)

Gambar 8. Grafik (a) *loss*; (b) *accuracy* LSTM

Tabel 3 menunjukkan hasil perbandingan dari beberapa metode yang dirancang dalam penelitian. Dari tabel ini, dapat dilihat bahwa jaringan Bi-LSTM menunjukkan keunggulan dalam kinerja secara keseluruhan. Rata-rata nilai *F1-Score* dan nilai ROC AUC yang dicapai dengan mempelajari fitur dalam bentuk MFCC lebih tinggi daripada jaringan lainnya. Dari gambar 6 dan gambar 8, juga dapat dilihat bahwa jaringan Bi-LSTM mencapai akurasi validasi tertinggi dengan jumlah *epoch* yang lebih sedikit dibandingkan dengan jaringan LSTM. Dengan kata lain, jaringan Bi-LSTM konvergen lebih cepat dibandingkan dengan jaringan LSTM dan CNN. Berdasarkan hasil tersebut, paper ini mengusulkan metode Bi-LSTM sebagai metode yang paling baik dalam memprediksi SER.

Tabel 3. Perbandingan Model

Model	*F1-Score*	ROC AUC
Bi-LSTM	0,878	0,979
LSTM	0,836	0,971
CNN	0,873	0,977

IV. KESIMPULAN

Berdasarkan hasil analisis dari setiap sampel suara didapatkan kesimpulan bahwa fitur penentu yang secara pasti dapat menjelaskan emosi seseorang melalui suara belum dapat ditentukan secara pasti.

Hal ini terjadi karena sifat natural manusia yang memiliki gaya berbicaranya sendiri dan juga intonasi serta pronounsiasi kata yang dapat mengganggu penentuan fitur. Akan tetapi, ada pola pada jeda, frekuensi suara, dan amplitudo yang dapat cukup menjelaskan emosi dari pembicara.

Dalam pembuatan model dengan *deep learning,* yaitu menggunakan metode Bi-LSTM yang dibandingkan dengan metode LSTM, dan CNN diperoleh hasil bahwa jaringan Bi-LSTM menunjukkan keunggulan dalam kinerja secara keseluruhan. Rata-rata nilai *F1-Score* dan nilai ROC AUC yang dicapai dengan mempelajari fitur dalam bentuk MFCC lebih tinggi daripada jaringan lainnya. Dengan demikian, jaringan Bi-LSTM konvergen lebih cepat dibandingkan dengan jaringan LSTM dan CNN. Oleh karena itu, metode Bi-LSTM yang diusulkan dalam paper ini dianggap sebagai metode yang paling baik dalam memprediksi SER. Pada penelitian selanjutnya dapat menggunakan data yang lebih heterogen dan melakukan *preprocessing* untuk pemilihan datanya sehingga tidak terlalu banyak *noise* yang ada. Pemilihan data yang heterogen sangat penting karena nyatanya dalam mengungkapkan emosi yang dirasakan, manusia mempunyai nada dan keras suara yang berbeda-beda.

REFERENSI

1. Azhari, "IMPLEMENTASI ALGORITMA CONVOLUTIONAL NEURAL NETWORK DALAM DETEKSI EMOSI MANUSIA BERDASARKAN EKSPRESI WAJAH," Jul. 08, 2021. http://eprosiding.ars.ac.id/index.php/pti/article/view/198.
2. Rere, LM Rasdi. "Studi Pengenalan Ekspresi Wajah Berbasis Convolutional Neural Network." Prosiding SeNTIK 3.1 (2019): 71-78.
3. S. Grover and A. Verma, "Design for emotion detection of punjabi text using hybrid approach," in Proceedings of the International Conference on Inventive Computation Technologies, ICICT 2016, 2017, vol. 2.
4. M. P. Skenduli, M. Biba, C. Loglisci, M. Ceci, and D. Malerba, "User-Emotion Detection Through Sentence-Based Classification Using Deep Learning: A Case-Study with Microblogs in Albanian Marjana," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11177 LNAI, Springer International Publishing, pp. v–vi, 2018.
5. A. A. Sundawa, A. G. Putrada, dan N. A. Suwastika, "Implementasi dan Analisis Simulasi Deteksi Emosi Melalui Pengenalan Suara Menggunakan Mel-Frequency Cepstrum Coefficient dan Hidden Markov Model Berbasis IOT."
6. B. Schuller, G. Rigoll, dan M. Lang, "SPEECH EMOTION RECOGNITION COMBINING ACOUSTIC FEATURES AND LINGUISTIC INFORMATION IN A HYBRID SUPPORT VECTOR MACHINE-BELIEF NETWORK ARCHITECTURE."
7. P. Juslin dan K. Scherer, "Speech emotion analysis," *Scholarpedia*, vol. 3, no. 10, hlm. 4240, 2008, doi: 10.4249/scholarpedia.4240.Author 1, A.; Author 2, B. *Book Title*, 3rd ed.; Publisher: Publisher Location, Country, 2008; pp. 154–196.
8. K. Wang, N. An, B. N. Li, Y. Zhang, dan L. Li, "Speech emotion recognition using Fourier parameters," *IEEE Trans Affect Comput*, vol. 6, no. 1, hlm. 69–75, Jan 2015, doi: 10.1109/TAFFC.2015.2392101.
9. F. KASYIDI, R. ILYAS, dan N. M. ANNISA, "Peningkatan Kemampuan Pengenalan Emosi Melalui Suara dalam Bahasa Indonesia," *MIND Journal*, vol. 6, no. 2, hlm. 194–204, Des 2021, doi: 10.26760/mindjournal.v6i2.194-204.
10. A. Graves, "Supervised Sequence Labelling," 2012, hlm. 5–13. doi: 10.1007/978-3-642-24797-2_2.
11. F. W. Wibowo dan Institute of Electrical and Electronics Engineers, *2018 International Conference on Information and Communications Technology (ICOIACT) : 6-7 March 2018*.
12. Yildirim Ö (2018) A novel wavelet sequences based on deep bidirectional LSTM network model for ECG signal classification. Comput Biol Med 96:189–202. https://doi.org/10.1016/j.compbiomed.2018.03.016.