JINTEKS (Jurnal Informatika Teknologi dan Sains) ISSN 2686-3359 (Online)
Vol. 4 No. 3, Agustus 2022, hlm. 220 – 224

DETEKSI EMOSI BERDASARKAN WICARA MENGGUNAKAN DEEP LEARNING MODEL

Siska Rahmadani[1*], Cicih Sri Rahayu[2], Agus Salim[3], Karno Nur Cahyo[4]
[1,2,3,4]Ilmu Komputer, Universitas Nusa Mandiri
email: 14002456@nusamandiri.ac.id[*]

Abstrak: Kemampuan komputer dalam meniru kemapuan manusia sampai saat ini menjadi hal yang menarik untuk dikembangkan. Di beberapa studi, pengenalan emosi sudah banyak diteliti baik melalui foto wajah dan ucapan verbal maupun non verbal. Penelitian ini bertujuan untuk mengeksplorasi berbagai metode *Deep Learning* untuk mendapatkan model yang paling baik dalam mendeteksi emosi menggunakan *dataset* EmoDB. *Feature extraction* dilakukan dengan menggunakan Zero Crossing Rate, Chroma_stft, Mel Frequency Cepstral Coefficients (MFCC), Root Mean Square (RMS) dan MelSpectogram. Pada tahap *preprocessing* diterapkan teknik data augmentation dengan mengaplikasikan *noise injection*, *shifting time* dan mengubah *pitch* dan kecepatan audio. Dari hasil penelitian dikemukakan bahwa metode deep learning terbaik berdasarkan nilai akurasi adalah CNN-BiLSTM.

Kata Kunci : Deteksi Emosi, EmoDB, *Deep Learning, Feature Extraction*

Abstract: *The ability of computers to imitate human abilities has been an interesting thing to develop. In several studies, emotion recognition has been studied both through facial photos and verbal and non-verbal speech. This study aims to explore various deep learning methods to get the best model for detecting emotions using the EmoDB dataset. Feature extraction is done using Zero Crossing Rate, Chroma_stft, Mel Frequency Cepstral Coefficients (MFCC), Root Mean Square (RMS) and MelSpectogram. In the pre-processing stage, data augmentation techniques are applied by applying noise injection, shifting time and changing the audio pitch and speed. From the results of the study, it was stated that the best deep learning method based on the accuracy value was CNN-BiLSTM.*

Keywords **:** Emotion Detection, EmoDB, *Deep Learning, Feature Extraction*

PENDAHULUAN

Seseorang dalam menyampaikan emosi dan perasaan dapat melalui ekspresi wajah dan ucapan. Seiring perkembangan teknologi, semakin banyak aplikasi terkait yang muncul seperti *personal digital assistant* (PDA), sensor, *text-to-speech* dan sebagainya yang menggunakan input suara atau ucapan. Deteksi emosi dari ucapan menjadi topik penelitian yang menarik sehubungan dengan bagaimana melatih komputer untuk dapat meniru keterampilan manusia. Dengan bantuan *emotion recognition* komputer juga dapat lebih baik dalam membuat keputusan yang dapat membantu penggunanya [1].

Deep Learning merupakan pendekatan yang banyak digunakan untuk pengenalan emosi. Hal ini karena pendekatan menggunakan metode tradisional dianggap lambat dan tidak akurat. Karena data yang cenderung besar, metode tradisional menyembabkan kelambatan proses sehingga menjadi mahal. Untuk bidang yang membutuhkan tingkat akurasi tinggi, seperti di bidang kesehatan metode tradisional menjadi tidak efektif. Oleh karena itu *model end to end* seperti *Deep Learning* lebih cocok digunakan karena menggunakan rangkaian operasi matematika (tensormath) yang sama dengan GPU sehingga memungkinkan mempercepat proses komputasi. *Deep Learning* dapat memberikan informasi yang lebih akurat tentang perilaku anomali karena dapat menunjukkan masalah utama dan asal intrusi, juga mampu mengevaluasi data besar dengan waktu yang lebih cepat dibandingkan dengan metode tradisional serta beradaptasi dengan perubahan konteks [2].

Deteksi emosi ataupun deteksi pengenalan ekspresi wajah menjadi topik yang saat ini banyak diteliti di area *Deep Learning* sebagai langkah besar menuju interaksi *Advanced Human Computer*. Oleh karena itu banyak metode-metode Deep Learning yang diusulkan oleh berbagai peniliti seperti metode untuk mendeteksi emosi berdasarkan wicara. Penelitian oleh [3] mengusulkan Tsception yaitu sebuah kerangka *Deep Learning* yang digunakan untuk mendeteksi emosi dari *electroencephalogram* (EEG) untuk dibandingkan dengan hasil dari model SVM, EEGNET dan LSTM. CNN sebagai salah satu metode *Deep Learning* juga banyak diaplikasikan untuk mendeteksi emosi seperti penelitian yang dilakukan oleh [4] dan [5].

Pada area medis, deteksi suara dapat digunakan untuk mendeteksi penyakit Parkinson [6] dengan menggunakan LSTM. Model LSTM juga bekerja dengan baik digunakan untuk mendeteki ganguan suara patologis (*pathological voice disorder*) [7]. Model *Deep Learning* GRU untuk deteksi suara nyanyi mampu menghasilkan performa yang baik pada Jamendo dan RWC *dataset* [8].

Saat ini masing-masing metode memiliki kelebihan dan kekurangan yang ada, pada penelitian ini berfokus pada mengeksplorasi lima metode deep learning yang dapat digunakan untuk mendeteksi emosi berdasarkan wicara dengan menggunakan beberapa transformasi untuk mengekstraksi feature yaitu Zero Crossing Rate, Chroma_stft, Mel

ISSN 2686-3359 (Online)

Frequency Cepstral Coefficients (MFCC), Root Mean Square (RMS) dan MelSpectogram. Pada penelitian ini lima metode CNN, BiLSTM, BiGRU, CNN-BiGRU dan CNN-BiLSTM akan dianalisis menggunakan EmoDB Dataset, untuk mendapatkan satu metode Deep Learning yang memiliki akurasi terbaik diantaranya.

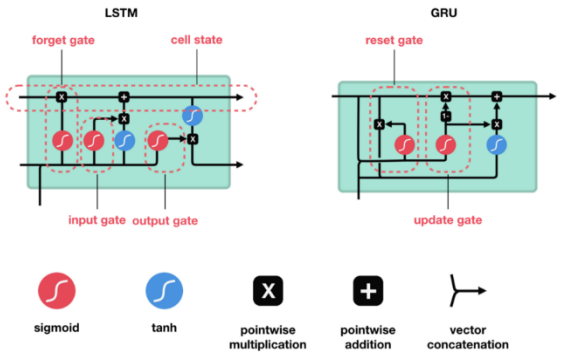
TINJAUAN PUSTAKA

Convolutional Neural Network (CNN)

Metode deep learning dikenal sukses digunakan dalam area image processing. CNN sebagai salah satu arsitektur dari *Deep Learning*, memiliki kemampuan yang baik pada berbagai tugas *visual recognition* [9]. Meskipun CNN dirancang untuk digunakan pada pengenalan objek dalam gambar, tetapi jika dibandingkan dengan algoritma klasik *machine learning* lainnya, CNN memiliki kemampuan yang lebih baik dalam mendeteksi emosi [10]. CNN dapat bekerja dengan baik apabila diberi banyak sinyal seperti perubahan frekuensi dan perubahan amplitudo yang ditemukan di dalam speech [11]. Kemampuan model CNN dapat ditingkatkan dengan dikombinasikan teknik *feature extraction* [4].

LSTM dan GRU

Long Short-Term Memory (LSTM) dan *Gated Recurrent Unit* (GRU) adalah model *Deep Learning* yang diciptakan untuk mengatasi masalah short term memory yang bisa muncul pada model *Recurrent Neural Network* (RNN). Jika urutan langkah yang melalui node cukup panjang, RNN akan kesulitan membawa informasi dari node sebelumnya ke langkah selanjutnya. RNN mungkin meninggalkan informasi penting sejak awal. Gambar 1 menggambarkan arsitektur dari LSTM dan GRU



Gambar 1. Arsitektur LSTM dan GRU

Model *Bidirectional* LSTM (BiLSTM) terdiri dari dua model LSTM dan *Bidirectional* GRU (BiGRU) terdiri dari dua model BRU, yang satu *forward direction* dan satu lagi *backward direction*. Proses *training* menggunakan BiLSTM maupun BiGRU mungkin saja lebih lama, tetapi memungkinkan model menangkap beberapa fitur tambahan yang tidak bisa ditangkap oleh model LSTM dan GRU [12].

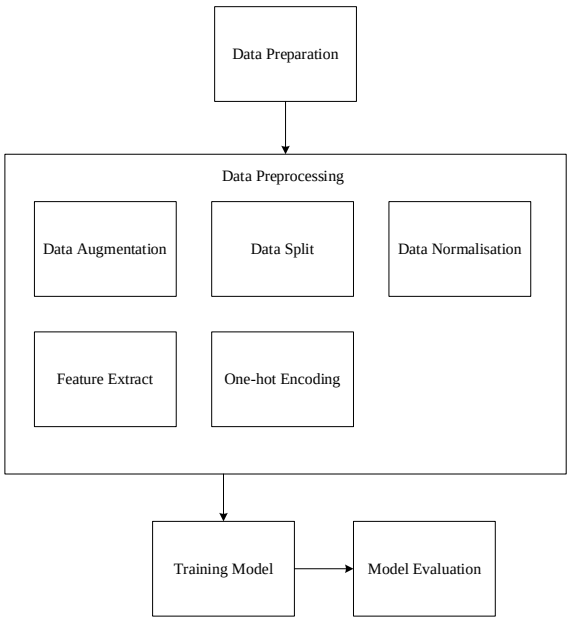
Feature Extraction

Ekstraksi Fitur dilakukan untuk mengurangi jumlah fitur dalam kumpulan data dengan membuat fitur baru dari yang sudah ada dan membuang fitur asli. Kumpulan fitur baru yang dikurangi ini merangkum sebagian besar informasi yang terkandung dalam kumpulan fitur asli sehingga versi ringkasan dari fitur asli dapat dibuat dari kombinasi set asli. Beberapa diantara metode *feature extraction* yang digunakan salah satunya adalah MFCC [13]

MFCC didasarkan pada persepsi pendengaran manusia yang tidak dapat merasakan frekuensi lebih dari 1Khz dan memiliki dua jenis filter yang berjarak linier pada frekuensi rendah di bawah 1000 Hz dan jarak logaritmik di atas 1000Hz [13].

METODE

Penelitian ini menggunakan dataset yang bersumber dari emoDB *Dataset*, terdiri dari basis data emosional Jerman yang dikumpulkan dan dikelola oleh *Institute of Communication Science, Technical University*, Berlin, Germany [14]. EmoDB *dataset* terdiri dari 535 ucapan yang berisi 7 emosi pada kepribadian manusia, emosi tersebut akan menjadi label pada *dataset* ini. 7 emosi tersebut adalah *anger* atau marah, *boredom* atau bosan, *anxiety* atau cemas, *happiness* atau bahagia, *sadness* atau sedih, *disgust* atau jijik, *neutral* atau bersifat netral.



Gambar 2. Tahapan Penelitian

Tahapan penelitian digambarkan pada gambar 2. Pada tahapan data preparation, untuk mendapatkan kategori emosi, nama file di *parsing* sesuai dengan *emotion code*. Data *augmentation* dilakukan dengan menambahkan sampel data sintaksis baru. Teknik ini juga dapat membantu

ISSN 2686-3359 (Online)

menghindari masalah *overfitting*. Agar dapat menghasilkan data sintaksis untuk audio, kita dapat menerapkan *noise injection, shifting time*, mengubah *pitch* dan kecepatan audio sehingga dapat meningkatkan kemampuan model dalam mengeneralisasi.

Feature extraction dilakukan karena data yang digunakan tidak terstruktur berupa audio atau suara, maka tidak dapat dipahami secara langsung oleh model algoritma. Data perlu diubah terlebih dahulu menjadi format yang dapat dipahami. Dengan *sample rate* dan *sample* data, sekarang kita dapat menggunakan beberapa metode *feature extraction* untuk melakukan transformasi data.

Karena pengklasifikasian menggunakan data dengan *multiclass*, maka diterapkan *one-hot encoding* pada *class*. *One-hot encoding* umum digunakan untuk menangani *multiclass classification task* karena keefektifan dan kesederhanaan [11]. Selanjutnya data di split menjadi *test set* dan *train set* dengan *ratio* 0,25:0,75 untuk masing-masing *test set* dan *train set*. Agar rentang nilai pada *dataset* tidak terlalu bervariasi, dan tetap berada dalam rentang atau *scale* yang sama, maka diterapkan normalisasi data menggunakan Sklearn *StandardScaler*

Training model menggunakan model deep learning yaitu CNN, BiLSTM, BiGRU, CNN-BiGRU dan CNN-BiLSTM. Model dievaluasi dan dibandingkan berdasarkan nilai akurasi masing-masing model

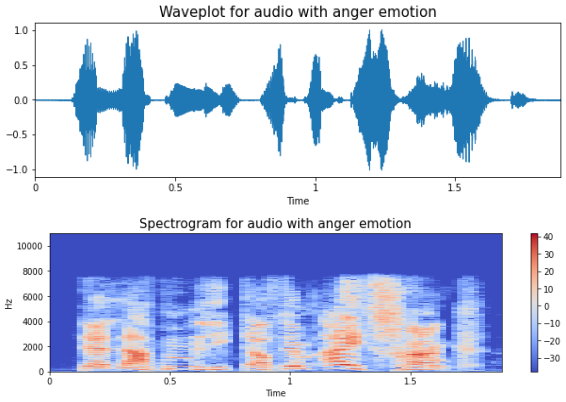
HASIL DAN PEMBAHASAN

Untuk mendapatkan jenis emosi (*emotion*) dari emoDB *Dataset*, maka nama *file* dicocokkan sesuai dengan *emotion code*, dengan rujukan pedoman pengkategorian tipe emosi. Semisal nama *file* tertulis 08b09Lc.wav, maka masuk dalam label *Boredom*, karena terdapat huruf L pada 2 karakter terakhir nama file yang dapat diartikan bahwa file tersebut berlabel Langeweile atau boredom. Setelah dilakukan pengkategorian, muncul distribusi data dari tujuh tipe emosi yang ada, dengan distribusi data *Anger* (marah) ini yang paling banyak dengan 127 data, sedangkan tipe emosi yang distribusi datanya paling rendah adalah data *disgust* merasa jijik, dengan 46 data, rincian distribusi data tujuh tipe emosi ditampilkan pada Tabel 1.

Tabel 1. Distribusi Tipe Emosi

Emotions	Count
Anger	127
Anxiety	69
Boredom	81
Disgust	46
Happiness	71

Selain mekanisme data *visualization* dalam bentuk distribusi data, dari data audio dapat dimunculkan format data *visualization* dalam bentuk *waveplot* dan juga *spectogram* menggunakan *feature* librosa. Contoh data *visualization* dalam bentuk *waveplot* dan *spectogram* untuk emosi *anger* (marah) ditampilkan pada Gambar 3.



Gambar 3. Data Visualization dalam bentuk WavePlot dan Spectogram

Proses selanjutnya menggunakan teknik data augmentation pada data sample audio, yang diterapkan untuk menghindari *overfitting*, dari sample audio setelah ditambahkan teknik data augmentation akan terpecah menjadi 4 jenis data audio yaitu *noise injection*, *shifting*, *stretching* dan juga *pitch*. Setelah melalui proses data augmentation selanjutnya adalah mengimplementasikan penerapan feature extraction, karena *dataset* berupa suara atau audio, maka data tersebut disebut *unstructured* data. *Unstructured* data tidak dapat dipahami secara langsung oleh model, maka *unstructured* data harus diubah menjadi format yang dapat dipahami. *Feature extraction* pada data ini menggunakan Zero Crossing Rate, Chroma_stft, Mel Frequency Cepstral Coefficients (MFCC), Root Mean Square (RMS) dan MelSpectogram. Hasil yang didapatkan setelah melakukan *feature extraction* adalah mendapatkan 127 *feature input* dan 1 *feature* label.

Training model menggunakan metode *deep learing*. Model yang digunakan adalah CNN, BiLSTM, BiGRU, CNN-BiLSTM dan CNN-BiGRU. Agar perbandingan model setara, maka nilai parameter *batch_size* dan *epoch* yang digunakan juga sama yaitu *batch_size* 64 dan *epoch* 50. Semua model menggunakan arsitektur dengan *optimizer* Adam dan *loss function* yaitu *categorical cross-entropy*.

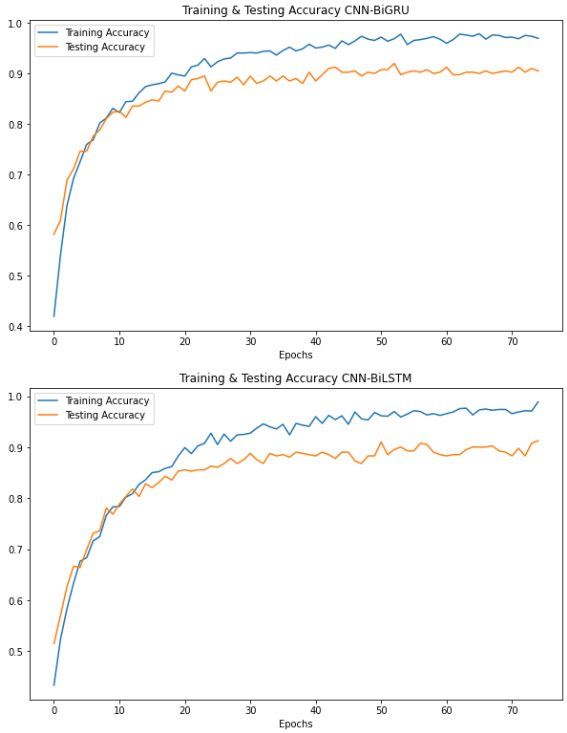
Arsitektur model CNN terdiri dari empat *convolutional layer* dengan *kernel size* 5x5, *strides* 1x1, *padding=same*, *hfilters*=50, relu *activation*, dan *maxpooling*, satu *fully connected layer* dengan jumlah node 32 dengan relu *activation* dan *ouput* node 7 dengan sofmax *activation*. Hasil yang didapatkan adalah nilai akurasi model CNN sebesar 81,34% akurasi.

Arsitektur model BiLSTM terdiri dari satu *layer* BiLSTM dengan input *node* 256 dan maxpooling. Terdapat juga satu *fully connected layer* dengan node 32 dan relu *activation*, *ouput node* 7

ISSN 2686-3359 (Online)

dengan sofmax activation. Dari hasil *training* model didapatkan nilai akurasi dari model BiLSTM adalah sebesar 88,05%. Model BiGRU mempunyai *input node*, *fully connected layer*, *output node* dan *activation function* yang sama dengan BiLSTM. Nilai akurasi yang didapatkan dari model BiGRU adalah 90,29%., lebih tinggi 2,24% dibandingkan dengan hasil pada model BiLSTM.

Pada penelitian ini juga dilakukan eksplorasi model CNN digabungkan dengan model BiLSTM dan BiGRU. Dengan menggunakan arsitektur satu *convolutional* layer dengan *kernel size* 3x3, *strides* 1x1, *padding=same*, *hfilters*=50, relu *activation*, *maxpooling*, *dropout* 0,5. Terdapat satu *layer Bidirectional* LSTM (pada model CNN-BiLSTM) atau *Bidirectional* GRU (pada model CNN-BiGRU) dengan *node* 128, *dropout* 0,5. Ditambah dengan satu *fully connected layer* dengan 128 *node*, *dropout* 0,5 dan *ouput layer* dengan *node* 7 dan *sofmax activation*. Hasil yang didapatkan adalah nilai akurasi sebesar 91,29% oleh model CNN-BiLSTM dan 90,54% oleh model CNN-BiGRU dengan perbandingan akurasi dapat dilihat pada gambar 4. Dimana dapat dilihat terdapat peningkantan nilai akurasi apabila dibandingkan dengan model CNN, BiLSTM dan BiGRU.



Gambar 4. Plot Training dan Testing Accuracy CNN-BiGRU dan CNN-BiLSTM

Dari hasil yang didapatkan akurasi tertinggi yaitu 91,29% didapatkan dengan menggabungkan model CNN dan BiLSTM. Hal ini memungkinkan dimana kemampuan CNN menangkap informasi spasial (konteks) dilengkapi oleh kemampuan BiLSTM sebagai RNN network dalam menangkap informasi sekuensial. Perbandingan hasil nilai akurasi masing-masing model selengkapkan dapat dilihat pada tabel 2.

Tabel 2. Perbandingan Hasil Akurasi Model

Model	Accuracy
CNN	81,34%
BiLSTM	88,05%
BiGRU	90,29%
CNN-BiLSTM	91,29%
CNN-BiGRU	90,54%

KESIMPULAN DAN SARAN

Penelitian deteksi emosi menggunakan lima model *Deep Learning* untuk pengolahan data audio atau suara, antara lain CNN, BiLSTM, BiGRU, CNN-BiLSTM dan CNN-BiGRU. Berdasarkan perbandingan hasil akurasi model yang muncul, model CNN-BiLSTM merupakan model dengan akurasi performa yang paling baik dalam mendeteksi emosi berdasarkan ucapan yaitu 91,29%. Sedangkan model yang akurasi performanya paling rendah dalam mendeteksi emosi berdasarkan ucapan adalah model LSTM dengan nilai 81,34%.

Saran penelitian selanjutnya bisa menerapkan metode *hyper parameter tuning* untuk meningkatkan performa semua model, metode *hyper parameter tuning* yang bisa diterapkan adalah metode GridSearchCV atau RandomSearchCV.

DAFTAR PUSTAKA

[1] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj, "Emotion detection from text and speech: a survey," *Soc. Netw. Anal. Min.*, vol. 8, no. 1, 2018.

[2] B. Dong and X. Wang, "Comparison deep learning method to traditional methods using for network intrusion detection," *Proc. 2016 8th IEEE Int. Conf. Commun. Softw. Networks, ICCSN 2016*, pp. 581–585, 2016.

[3] Y. Ding *et al.*, "TSception:A Deep Learning Framework for Emotion Detection Using EEG," *Proc. Int. Jt. Conf. Neural Networks*, 2020.

[4] S. Kwon, "A CNN-Assisted Enhanced Audio Signal Processing," *Sensors*, 2020.

[5] A. Jaiswal, A. Krishnama Raju, and S. Deb, "Facial emotion detection using deep learning," *2020 Int. Conf. Emerg. Technol. INCET 2020*, pp. 1–5, 2020.

[6] D. R. Rizvi, I. Nissar, S. Masood, M. Ahmed, and F. Ahmad, "An LSTM based deep learning model for voice-based detection of Parkinson's disease," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 5 Special Issue, pp. 337–343, 2020.

[7] V. Gupta, "Voice Disorder Detection Using Long Short Term Memory (LSTM) Model," 2018.

[8] R. Monir, D. Kostrzewa, and D. Mrozek, "Singing Voice Detection: A Survey," *Entropy*, vol. 24, no. 1, 2022.

[9] A. Shewalkar, D. nyavanandi, and S. A. Ludwig, "Performance Evaluation of Deep neural networks Applied to Speech Recognition: Rnn, LSTM and

 ISSN 2686-3359 (Online)

GRU," *J. Artif. Intell. Soft Comput. Res.*, vol. 9, no. 4, pp. 235–245, 2019.

[10] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, and N. Arunkumar, "Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS)," *IEEE Access*, vol. 7, pp. 57–67, 2019.

[11] D. Nagajyothi and P. Siddaiah, "Speech recognition using convolutional neural networks," *Int. J. Eng. Technol.*, vol. 7, no. 4.6 Special Issue 6, pp. 133–137, 2018.

[12] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The Performance of LSTM and BiLSTM in Forecasting Time Series," *Proc. - 2019 IEEE Int. Conf. Big Data, Big Data 2019*, pp. 3285–3292, 2019.

[13] L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," vol. 2, no. 3, pp. 138–143, 2010.

[14] "Emo-DB." [Online]. Available: http://emodb.bilderbar.info/download/. [Accessed: 05-Aug-2022].