ANALISA ALGORITMA CONVOLUTION NEURAL NETWORK (CNN) PADA KLASIFIKASI GENRE MUSIK BERDASAR DURASI WAKTU

Yisti Vita Via, Intan Yuniar Purbasari, Aditya Putra Pratama
Universitas Pembangunan Nasional Veteran Jawa Timur
Email: yistivia.if@upnjatim.ac.id

Abstrak. *Genre musik pada umumnya digolongkan berdasarkan kemiripan ritmik, frekuensi dan harmoni. Biasanya masyarakat memilih genre musik sesuai kesenangan mereka. Pada penelitian ini, algoritma Convolutional Neural Network (CNN) digunakan untuk klasifikasi genre musik. Ekstraksi fitur yang digunakan antara lain: Chroma stft, Spectral centroid, Spectral bandwidth, Spectral Rolloff , Root Mean Square Energy (RMSE), Zero_Crossing Rate, Mel-Frequency Cepstral Coefficients, Harmony, Tempo, dan Perceptron. Perbedaan durasi waktu musik pada data uji coba adalah 10 detik dan 30 detik. Hasil uji coba pada data durasi 10 detik menghasilkan akurasi prediksi yaitu sebesar 81%. Sedangkan hasil uji coba pada data durasi 30 detik menghasilkan akurasi prediksi yaitu sebesar 58%. Hal ini dapat disimpulkan bahwa klasifikasi genre musik dengan durasi waktu yang lebih pendek ternyata mampu menghasilkan nilai akurasi yang lebih baik.*

Kata Kunci: Genre, Musik, Klasifikasi, CNN

Pada saat ini teknologi informasi banyak digunakan untuk mempermudah pekerjaan manusia di berbagai bidang, salah satunya musik. Genre musik merupakan irama yang digolongkan berdasarkan kemiripan ritmik, frekuensi, dan harmoni. Masyarakat biasanya mengelompokkan genre musik berdasarkan kesenangan mereka. Menggolongkan genre musik bisa dilakukan dengan mendengarkan file musik secara langsung. Namun jika jumlah musik yang akan dikelompokkan itu sangat besar, jutaan mungkin, maka hal ini akan menjadi kendala. Selain terkendala waktu, hasil yang akurat dan presisi pun tidak mudah untuk dicapai.

Klasifikasi genre musik secara otomatis dapat membantu menyelesaikan permasalahan ini. Pengolahan sinyal digital pada sinyal audio telah berkembang pesat untuk mendukung terciptanya suatu sistem yang bekerja secara digital. Selanjutnya sangat diperlukan adanya metode dan algoritma untuk melakukan klasifikasi secara cepat dan akurat. Penelitian terkait klasifikasi genre musik ini telah banyak dilakukan, antara lain [1], [2], [3], [4], [5], [6], dan [7]. Topik ini memiliki peranan penting dalam beberapa aplikasi pemutaran musik. Salah satu peranan ini semisal pentingnya memberikan rekomendasi dalam pemilihan genre musik sesuai kenginan pengguna. Pada perkembangan selanjutnya dapat diarahkan pada pengecekan plagiarisme musik pada kesamaan nada, baik itu ritmik, frekuensi, ataupun harmoni.

Pada penelitian ini, awalnya file musik harus diekstraksi dahulu fiturnya dengan menggunakan metode Mel Frequency Cepstrum Coefficient (MFCC). MFCC bekerja dengan mengadopsi fitur kerja telinga, yaitu dengan membedakan tinggi rendah sinyal suara menggunakan speak recognition untuk mengetahui ciri dari setiap suara. Selanjutnya hasil ekstraksi fitur akan diklasifikasikan menggunakan Convolutional Neural network (CNN).

Sebagaimana penelitian yang berkembang bahwa CNN merupakan neural network yang biasa digunakan pada data gambar atau citra. Namun pada penelitian ini, CNN digunakan untuk klasifikasi genre musik berdasar fitur-fiturnya. Alasan utama penggunaan CNN adalah dikarenakan algoritma ini menggunakan dimensi lebih dari satu yang dapat mempengaruhi keseluruhan skala dalam objek. Hal ini sangat penting yang bertujuan agar input tidak kehilangan informasi spasial yang nantinya akan diekstraksi fitur dan diklasifikasi. Selain itu, hal ini juga akan mampu meningkatkan akurasi hasil klasifikasinya.

Pada penelitian sebelumnya, klasifikasi genre musik menggunakan CNN sudah banyak dikembangkan. Beberapa penelitian terkait topik tersebut antara lain [8], [9], dan [10]. Pada penelitian ini data set yang digunakan diambil dari GTZAN yang sangat terkenal di *Musik Information Retrieval* (MIR) [11].

Pada setiap data set audionya sudah dilengkapi gambar spektrum dengan durasi

waktu yang sama. Nantinya durasi ini akan dibuat menjadi dua variasi untuk mendukung kebutuhan analisa pengujian. Dengan peningkatan jumlah data set, jenis genre musik, serta variasi durasi waktu musik dalam penelitian ini, diharapkan mampu memberikan kontribusi analisa solusi terhadap performansi CNN yang lebih baik dalam kasus klasifikasi.

I. Metodologi

Pengumpulan dan Praproses Data

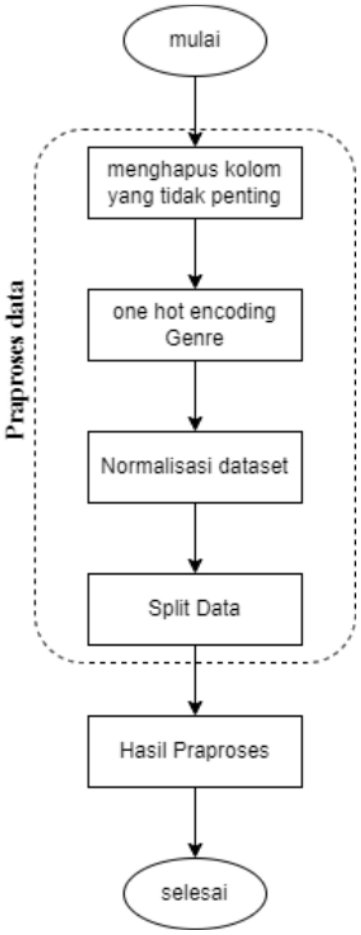
Sebagaimana dijelaskan di awal bahwa data set yang digunakan berjumlah 1000 file yang didapatkan dari GTZAN - *Music Genre Classification Data Set.* Data berupa file audio musik dengan tipe .wav berdurasi 30 detik. Total data set terbagi masing-masing dalam 100 file dalam setiap genre musik klasik, blues, country, hip-hop, disco, jazz, pop, metal, rock, dan reggae.

Keseluruhan data set file audio ini selanjutnya akan dilakukan praproses dengan memecah atau split lagu ke dalam 2 jenis durasi yang berbeda yaitu 10 detik dan 30 detik. Karena data awal berdurasi 30 detik, maka diperlukan proses pemecahan menjadi durasi 10 detik. Tahapan praproses data digambarkan pada Gambar 1. Dengan cara ini maka jumlah data set akan bertambah banyak. Semakin banyak data set yang digunakan nantinya akan cukup digunakan untuk meningkatkan performansi hasil klasifikasi pada pengujiannya.

Ekstraksi Fitur

Pada tahap ini file audio akan diproses dengan fitur-fitur audio yang ada di Phyton. Setiap file nantinya memiliki nilai Mean dan Variance yang dihitung dari beberapa fitur hasil ekstraksi setiap file audio. Hasil ekstraksi fitur data set tersimpan dalam file csv. Adapun fitur ekstraksi audio yang digunakan dalam data set ini antara lain:

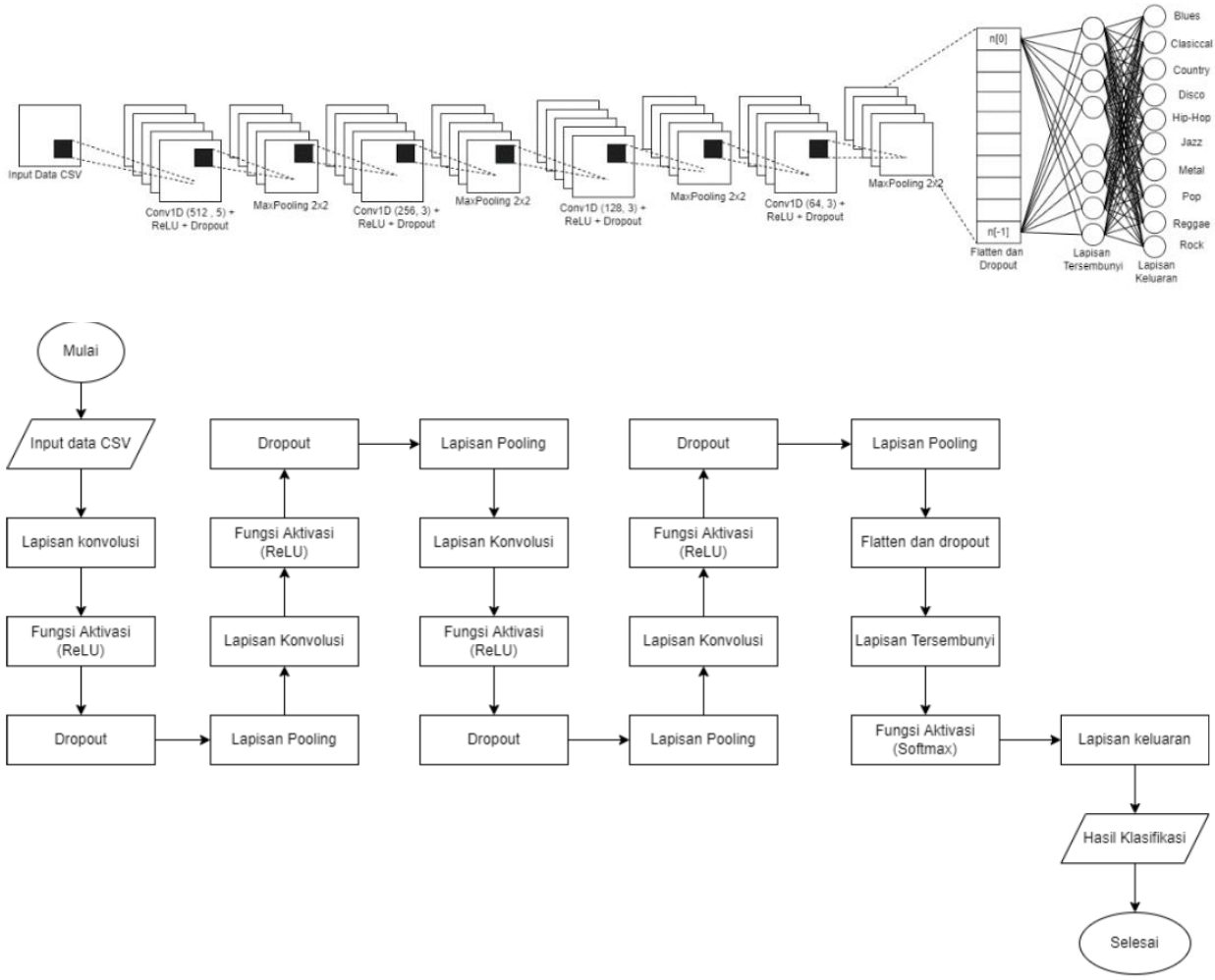
a. Chroma stft dengan menghitung kromagram dari bentuk gelombang atau spektogram daya.
b. Root-mean-square (RMS) dengan menghitung nilai root-mean-square (RMS) untuk setiap frame, baik dari sampel audio yatau dari spektogram.
c. Spectral centroid dengan menghitung centroid 26 pectral.
d. Spectral bandwidth dengan mengitung bandwidth spektral
e. Spectral rolloff dengan menghitung frekuensi roll-off.
f. Zero crossing rate dengan menghitung tingkat zero-crossing dari deret waktu audio.
g. Harmony dengan menghitung tempogram dari audio
h. Perceptron dengan mengekstrak elemen perkusi dari rangkaian waktu audio.
i. Tempo dengan perkiraan tempo detak per menit
j. Mel-Frequency Cepstral Coefficient dengan memproses audio dalam satu batch



Gambar 1. Tahapan Praproses Data

Arsitektur CNN

Arsitektur CNN dimulai dari masukan data csv hingga proses klasifikasi yang terdiri dari 4 lapisan convolusi dan 4 lapisan *pooling*. Pada Gambar 2 diilustrasikan arsiteksur keseluruhan dari proses CNN.



Gambar 2. Arsitektur CNN

Alur arsitektur CNN dimulai dengan memproses inputan csv yang berisi hasil fitur ekstraksi dari data audio pada tahap ekstraksi fitur. Selanjutnya tahapan dilanjutkan ke lapisan konvolusi dan ke fungsi aktivasi ReLU. Kemudian proses dilanjutkan ke lapisan max *pooling* dan proses *dropout*. Saat berada di proses *dropout* data akan diubah menjadi 1 dimensi karena akan masuk ke lapisan tersembunyi. Selanjutnya nilai dari data audio akan masuk ke fungsi aktivasi *softmax*, dan setelah diaktivasi maka akan dilanjutkan ke lapisan keluaran yang menghasilkan klasifikasi data.

Pelatihan dan Pengujian Model CNN

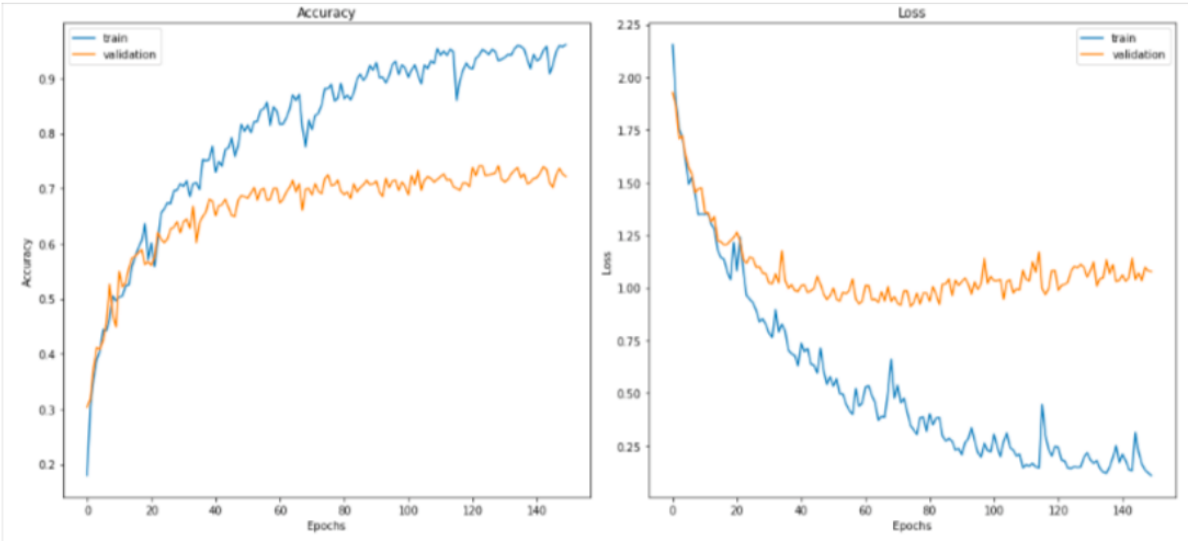
Pada tahap pelatihan model, akan digunakan data *training* dari csv. Data *training* diproses pada algoritma CNN dengan validasi dari data *training* itu sendiri. Hasil dari pelatihan ini nantinya berupa model jaringan yang sudah dilatih dan diuji dengan data *training*.

Model jaringan ini selanjutnya akan digunakan untuk tahap pengujian. Pada tahap ini digunakan data set validasi selain dari data *training*. Hasil pengujian akan menghasilkan nilai akurasi performansi klasifikasi genre musik yang nantinya dievaluasi menggunakan confusion matrix untuk mengetahui hasil performa dari model jaringan CNN yang sudah diperoleh pada pelatihan. Hasil confusion matrix digunakan untuk menghitung nilai akurasi, presisi, recall dan f1-score (10 Ghoneim 2019).

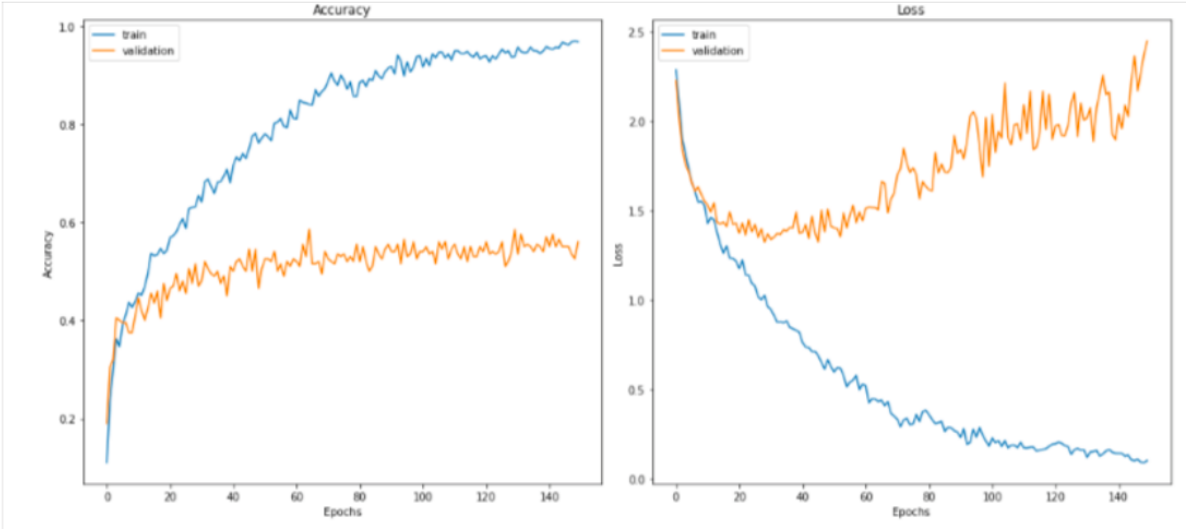
II. Hasil dan Pembahasan

Hasil Percobaan

Pada penelitian ini implementasi sistem menggunakan perangkat Keras yang ada di virtual machine yaitu Google Collab Pro dengan spesifikasi komputer adalah Intel Core i3 2.30 GHz, Ram DDR3 8GB.



Gambar 3. Hasil Pengujian Data Set File Audio Durasi 10 Detik



Gambar 4. Hasil Pengujian Data Set File Audio Durasi 30 Detik

Model jaringan CNN dilatih dengan GPU yang ada di Google Collab Pro. Perangkat lunak yang digunakan adalah Python 3, dengan library numpy, pandas, matplotlib, os, csv, librosa, dan bash. Sedangkan modul pendukung utama dalam klasifikasi genre musik ini adalah Tensorflow dan Librosa.

Terdapat dua skenario percobaan yang dilakukan dalam penelitian ini yaitu pengujian menggunakan data set file audio dalam durasi 10 detik dan 30 detik. Masing-masing hasil pengujian digambarkan grafik nilai akurasi dan *loss* terhadap jumlah epoch yang digunakan sebanyak 150. Pengujian dilakukan dengan menggunakan model CNN lapisan konvolusi berjumlah 4, dengan kernel 512 berjumlah 5, kernel 256 berjumlah 3, kernel 64 berjumlah 3, dan kernel 32 berjumlah 3.

Gambar 3 merupakan grafik dari skenario pengujian yang pertama yaitu dengan menggunakan dataset hasil split audio durasi 10 detik. Pada skenario ini menghasilkan nilai loss 0.0813, nilai akurasi 0.9709, nilai val_loss 0.8516, dan nilai val_accuracy 0.8098.

Sedangkan pada Gambar 4 adalah merupakan grafik dari skenario pengujian dengan menggunakan dataset hasil split audio durasi 30 detik. Pada skenario ini menghasilkan nilai loss 0.0304, nilai akurasi 0.9887, nilai val_loss 2.5737, nilai val_accuracy 0.5850.

Analisa dan Evaluasi Confusion Matrix

Selain pengukuran performansi pengujian dengan variabel loss, nilai akurasi, nilai val_loss, dan nilai val_accuracy, pengujian terhadap dua skenario juga diukur dari hasil confusion matriks dan heatmap. Hasil

confusion matriks pada pengujian skenario yang pertama yaitu menggunakan dataset hasil split audio dengan durasi 10 detik, disajikan pada Tabel 1. Sedangkan hasil pengujian hetmapnya terangkum bahwa akurasi per genre yaitu blues 77%, classical 94%, country 69%, disco 81%, hiphop 81%, jazz 81%, metal 95%, pop 84%, reggae 80%, dan rock 89%. Dari data ini disimpulkan bahwa akurasi terendah pada genre country dan akurasi tertinggi adalah genre metal.

Selanjutnya untuk confusion matriks skenario pengujian kedua yaitu yang menggunakan dataset hasil split audia dengan durasi 30 detik, disajikan pada Tabel 2. Hasil pengujian hetmapnya terangkum bahwa akurasi per genre yaitu blues 45%, classical 80%, country 42%, disco 48%, hiphop 53%, jazz 63%, metal 83%, pop 73%, reggae 46%, dan rock 56%. Dan ternyata disimpulkan bahwa akurasi terendah pada genre country dan akurasi tertinggi adalah genre metal, nilai ini dama dengan nilai pada hasil skenario pertama yang menggunakan durasi 10 detik.

Tabel 1. Confusion Matriks pada Dataset Durasi 10 Detik

	Blues	Classical	Country	Disco	Hiphop	Jazz	Metal	Pop	Reggae	Rock
Blues	36	0	5	0	2	0	2	0	2	0
Classical	0	63	0	0	0	4	0	0	0	0
Country	1	1	43	2	1	2	0	2	4	6
Discox	1	0	1	54	2	0	0	2	3	4
Hiphop	0	0	1	1	50	0	1	3	3	3
Jazz	1	3	2	1	1	46	0	1	0	2
Metal	0	0	0	2	1	0	59	0	0	0
Pop	0	0	0	2	1	1	0	47	3	1
Reggae	1	0	0	2	2	0	1	1	44	3
Rock	0	1	1	1	2	1	5	1	2	52

Tabel 2. Confusion Matriks pada Dataset Durasi 30 Detik

	Blues	Classical	Country	Disco	Hiphop	Jazz	Metal	Pop	Reggae	Rock
Blues	10	0	3	3	3	0	2	0	0	1
Classical	0	20	0	0	1	4	0	0	0	0
Country	4	0	8	1	1	1	1	2	0	1
Disco	1	0	1	11	0	0	1	1	2	6
Hiphop	1	0	0	1	9	0	0	1	1	4
Jazz	3	1	0	0	1	12	0	0	0	2
Metal	0	0	1	1	0	0	15	0	0	1
Pop	0	0	2	1	1	0	0	11	0	0
Reggae	2	0	1	2	2	0	0	3	11	3
Rock	0	0	3	2	0	2	0	0	1	10

Hasil confusion matriks ini kemudian dievaluasi dengan menghitung nilai presisi, recall, dan F1-score. Evaluasi pada skenario pertama disajikan pada Tabel 3, sedangkan evaluasi pada skenario kedua disajikan pada Tabel 4.

Dari perbandingan pada Tabel 3 dan 4 dapat dilihat bahwa tingkat akurasi yang lebih baik terletak pada skenario yang pertama yaitu sebesar 81%, dibandingkan skenario kedua yang hanya 58%. Pada skenario kedua, durasi waktu lebih lama yaitu sebesar 30 detik sehingga kemungkinan terjadinya overfitting lebih besar dibandingkan skenario kedua yang menggunakan dataset dengan durasi waktu lebih pendek yaitu 10 detik.

Berdasarkan hasil percobaan yang sudah dilakukan dapat dipertimbangkan bahwa selain melakukan split data dengan durasi waktu yang lebih pendek, beberapa faktor lainnya juga dapat mempengaruhi tingkat akurasi, diantaranya adalah jumlah dataset untuk pelatihan yang lebih banyak, perolehan model CNN yang lebih tepat, serta pemilihan teknik ektraksi fitur yang benar.

Tabel 3. Evaluasi Pengujian dengan Dataset Durasi 10 detik

Genre	Presisi	Recall	F1-score	Akurasi
Blues	0.90	0.77	0.83	0.81
Classical	0.93	0.94	0.93	
Country	0.69	0.69	0.69	
Disco	0.82	0.81	0.81	
Hiphop	0.79	0.81	0.80	
Jazz	0.85	0.81	0.83	
Metal	0.87	0.95	0.91	
Pop	0.82	0.84	0.83	
Reggae	0.72	0.80	0.76	
Rock	0.73	0.69	0.71	

Tabel 4. Evaluasi Pengujian dengan Dataset Durasi 30 detik

Genre	Presisi	Recall	F1-score	Akurasi
Blues	0.48	0.45	0.47	0.58
Classical	0.95	0.80	0.87	
Country	0.42	0.42	0.42	
Disco	0.50	0.48	0.49	
Hiphop	0.50	0.53	0.51	
Jazz	0.63	0.63	0.63	
Metal	0.79	0.83	0.81	
Pop	0.61	0.73	0.67	
Reggae	0.73	0.46	0.56	
Rock	0.36	0.56	0.43	

III. Kesimpulan

Berdasarkan percobaan yang telah dilakukan, terbukti penelitian ini telah berhasil menyelesaikan klasifikasi genre musik dengan menerapkan algoritma CNN. Dua jenis data set diujikan pada model skenario yang sama, yaitu jumlah epoch sebanyak 150, lapisan konvolusi berjumlah 4, kernel 512 berjumlah 5, kernel 256 berjumlah 3, kernel 64 berjumlah 3, dan kernel 32 berjumlah 3.

Dua skenario yang diujikan yaitu percobaan menggunakan data set dengan durasi 10 detik dan 30 detik. Hasil dari pengujian itu masing-masing memberikan nilai loss, nilai akurasi, nilai val_loss, nilai val_accuracy, heatmap, confusion matriks, nilai presisi, recall, dan F1-score.

Dari hasil percobaan diperoleh hasil secara keseluruhan bahwa pengujian dengan data set durasi waktu 10 detik memperoleh hasil akurasi yang lebih baik dari pada pengujian dengan data set durasi waktu 30 detik. Beberapa faktor yang mempengaruhi hasil akurasi antara lain durasi waktu file audio, jumlah dataset untuk pelatihan, parameter pemodelan CNN, serta teknik ektraksi fitur.

Sebagai pengembangan dari penelitian ini selanjutnya, dapat dilakukan perubahan parameter pada model CNN dengan nilai-nilai yang lain seperti jumlah lapisan konvolusi, batch_size, jumlah epochs, dan dropout. Selain itu bisa dilakukan penambahan pada jumlah kelas genre dan ektraksi fitur audionya sehingga mesin dapat mengenali ciri genre audio lebih banyak lagi.

IV. Daftar Pustaka

[1] G. Tzanetakis, P. Cook (2002). Musical genre classification of audio signals, IEEE Trans. Speech Audio Process. 10 (5). 293–302.

[2] G. Marques, T. Langlois, F. Gouyon, M. Lopes, M. Sordo. (2011). Short-term feature space and music genre classification, J. New Music Res. 40 127–137.

[3] C.M. Yeh, L. Su, Y. Yang. (2013). Dual-layer bag-of-frames model for music genre classification, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 246–250.

[4] S. Shin, H. Yun, W. Jang, H. Park. (2019). Extraction of acoustic features based on auditory spike code and its application to music genre classification, IET Signal Process. 13 (2) 230–234.

[5] Y. Yu, S. Luo, S. Liu, H. Qiao, Y. Liu, L. Feng, Deep attention based music genre classification, Neurocomputing.

[6] T. Kobayashi, A. Kubota, Y. Suzuki. (2018). Audio feature extraction based on sub-band signal correlations for music genre classification, in: 2018 IEEE International Symposium on Multimedia, ISM, 2018, pp. 180–181.

[7] Y. Panagakis, C. Kotropoulos, G.R. Arce. (2009). Music genre classification via sparse representations of auditory temporal modulations, in: 2009 17th European Signal Processing Conference, pp. 1–5.

[8] Y.M. Costa, L.S. Oliveira, C.N. Silla. (2017). An evaluation of convolutional neural networks for music classification using spectrograms, Appl. Soft Comput. 52 28–38.

[9] H. Yang, W.-Q. Zhang. (2019). Music genre classification using duplicated convolutional layers in neural networks, in: INTERSPEECH 2019.

[10] C. Senac, T. Pellegrini, F. Mouret, J. Pinquier (2017). Music feature maps with convolutional neural networks for music genre classification, in: Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing, CBMI '17, ACM, New York, NY, USA, pp. 19:1–19:5.

[11] B.L. Sturm, The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use, arXiv preprint arXiv:1306.1461.