PENGGOLONGAN LAGU BERDASARKAN SPEKTOGRAM DENGAN *CONVOLUTION NEURAL NETWORK*

Albert Parlys[*)], Ajub Ajulian Zahra, dan Achmad Hidayatno

Departemen Teknik Elektro, Universitas Diponegoro
Jl. Prof. Sudharto, SH, Kampus UNDIP Tembalang, Semarang 50275, Indonesia

[*)]*E-mail: albertparlysresearch @gmail.com*

Abstrak

Saat ini terdapat banyak lagu yang sudah diproduksi di dunia. Lagu-lagu tersebut digolongkan ke dalam genre berbeda. Ada berbagai macam genre mulai dari pop, rock, classic, reggae, dubstep, dan lain-lain. Perbedaan genre ini disebabkan adanya ketidaksamaan melodi, ketukan, intonasi, dan ekspresi pada masing-masing genre. Saat ini terdapat banyak metode yang digunakan untuk mengenali sebuah lagu, di antaranya audioprint, penggolongan genre, pengenalan ketukan lagu, pengenalan lirik lagu, dan lain-lain. Metode yang dipakai selama ini menggunakan database dengan ciri dari jutaan lagu. Salah satu metode lain adalah dengan mengembangkan sistem identifikasi lagu dengan suatu jaringan saraf terlatih. Penelitian ini akan membahas perancangan sebuah sistem untuk menggolongan lagu berdasarkan spektogram. Masukan sistem berupa lagu dengan format audio MP3 yang diubah ke dalam bentuk spektogram kemudian dilatih menggunakan Convolutional Neural Network. Ciri lagu akan diperoleh kemudian diklasifikan ke dalam lima genre berbeda yaitu pop, rock, classic, dubstep, dan reggae. Berdasarkan hasil pelatihan dan pengujian dengan filter 3x3 didapat nilai akurasi penggolongan lagu sebesar 100% pada 750 data latih dan 98% pada 50 lagu data uji. Algoritme pembelajaran terbaik pada pelatihan dengan filter yang sama adalah algoritme Adam yang lebih cepat dibandingkan dengan Adadelta, Adagrad, dan SGD.

*Kata kunci**:** CNN, spektogram, Adam, Adagrad, Adadelta, SGD.*

Abstract

Currently there are many songs that have been produced in the world. The songs are classified into different genres. There are various genres ranging from pop, rock, classical, reggae, dubstep, and others. This genre difference occurs in melodic inequality, tapping, intonation, and expression in each genre. Currently there are many methods used to classify a song, such as audioprint, genre classification, song recognition, song lyrics, etc. The method used so far uses a database with features from millions of songs. The other best method is to create a song classification system with a neural network system. This research will discuss the design of a system for classifying songs based on spectrograph. By using spectrograph of songs to be used for training using Convolutional Neural Network. Characteristics of songs will be classified into five genres of pop, rock, classic, dubstep, and reggae. Based on the results of training and testing with 3x3 filter obtained classification for 750 training data by 100% and 50 songs of test data by 98%. The best learning algorithm in training with the same filter is Adam algorithm which is faster compared to Adadelta, Adagrad, and SGD.

Keywords: CNN, spectrograph, Adam, Adagrad, Adadelta, SGD

1. Pendahuluan

Lagu adalah ragam suara yang berirama. Setiap lagu yang dibuat memiliki ciri masing-masing sehingga membedakan antara satu dengan lainnya. Ciri tersebut dapat digunakan sebagai pengenal atau identitas lagu. Konsep identifikasi adalah mengenali sesuatu dari komponen yang dimilikinya contohnya lagu, dari irama contohnya *rock* dan *classic*, intonasi keras atau lembutnya suara penyanyi, lirik lagu, dan *chord* yang dimainkan, atau dari kombinasi keempatnya.

Akustik adalah ilmu yang mempelajari ciri-ciri penyusun suatu suara secara fisis dan matematis[1]. Terdapat beberapa ciri yang sering digunakan dalam pengenalan akustik. Beberapa diantaranya adalah pengenalan tekanan akustik, dan pengenalan frekuensi. Pengenalan frekuensi merupakan suatu pengenalan pola (*pattern recognition*) yang khusus untuk kasus suara.

Jaringan Saraf Tiruan (JST) adalah metode pembelajaran mesin yang menirukan bagaimana cara manusia dapat belajar mengenai hal-hal baru. Metode pembelajaran JST dibagi menjadi pembelajaran terbimbing (*supervised)* dan

tidak terbimbing (*unsupervised*)[2]. Metode terbaru perkembangan dari JST adalah *Convolutional Neural Network* (CNN) yang mana memiliki struktur 3 dimensi (3D) untuk mempelajari suatu pola.

Penelitian tentang penggolongan lagu sudah dilakukan oleh Yandre M.G. Costa, dkk pada jurnal yang berjudul *An evaluation of Convolutional Neural Networks for music classification using spectrograms*[3], penelitian ini menggunakan spektogram sebagai ekstrasi ciri untuk menggolongkan genre lagu menggunakan *Latin Music Database*, ISMIR 2004, dan *African Music Database* dengan keakuratan rata-rata sebesar 92%. Penelitian mengenai metode pembelajaran menggunakan CNN sudah pernah dilakukan oleh Dimtri Palaz, dkk tahun 2015 pada penelitian berjudul *Analysis of CNN-based Speech Recognition System using Raw Speech as Input*[4] dan *Convolutional Neural Networks-based Continous Speech Recognition Using Raw Speech Signal*[5] dengan masing-masing akurasi pengenalan suara adalah 97,3% dan 93,6%.

Penelitian ini merancang sebuah sistem untuk menggolongkan sebuah lagu. Masukan yang digunakan adalah sebuah lagu dalam format audio MP3 (.mp3). Proses pembelajaran menggunakan CNN. Metode ini memutuskan apakah sebuah lagu tersebut dapat diklasifikasikan ke dalam salah satu genre.

2. Metode

2.1. Algoritme dan Diagram Sistem

Sistem dirancang untuk dapat mengklasifikasikan genre suatu lagu berdasakan spektogram menggunakan CNN. Lagu-lagu dalam basis data akan diklasifikasikan ke dalam 5 genre yaitu classic, dubstep, pop, reggae, dan rock Sistem pengklasifikasian lagu dibagi ke dalam tiga modul utama, yaitu pengolah suara, pengolah citra, serta pembelajaran dan pengklasifikasian dengan CNN. Berikut adalah garis besar proses yang dilakukan dalam sistem penggolongan genre ini :

1) Mengambil data suara dalam basis data untuk diubah ke dalam format WAV.
2) Membuat spektogram dari data suara.
3) Melakukan pemotongan spektogram menjadi dimensi yang lebih kecil yaitu 128x128 pixel:
4) Pelatihan atau Pengujian
5) Jika pelatihan dilakukan, maka bobot CNN didapat. Jika pengujian dilakukan, maka penggolongan genre didapat

2.2. Pengolah Suara

Proses pengolah suara adalah dengan mengubah format dari lagu pada basis data yaitu berformat MP3 menjadi format WAV. Kemudian data suara dalam format WAV akan diubah ke dalam bentuk spektogram menggunakan aplikasi SoX dengan parameter sebagai berikut

Tabel 1. Parameter Spektogram

Parameter	Keterangan
Vertical Pixel density	200ppc
Horizontal Pixel density	50pps
Legend	No
Colour	Yes

Setiap pixel horizontal akan mewakili 20 ms sebagaiman standar yang digunakan pada pencuplikan suara (20-25ms/cuplik). Setiap pixel vertikal pada spektogram mewakili frekuensi pada waktu tersebut dalam satuan dBFS (*decibel below full scale*). Jika frekuensi *tuning* A_4 = 440 Hz, maka nilai A_7 = 7040 Hz. Setiap perpindahan not (A-B atau C-D) pada tiga oktaf tersebut (18 kali perpindahan) akan menghasilkan nilai frekuensi sebesar 366,6 Hz.. Untuk mewakili nilai jarak 1 not maka dibutuhkan pencuplikan lebih besar dari 1-pixel untuk tiap 366.6 Hz. Dengan menggunakan 128-pixel untuk mewakili suara dengan rentang suara 0-22,1 KHz maka akan didapat nilai frekuensi per pixel sebesar 172,6 Hz/pixel. Sesuai dengan nilai frekuensi/pixel dapat disimpulkan bahwa pengunaan 128-pixel vertikal sudah memenuhi standar pencuplikan Nyquist yaitu, $fs \geq 2 \times f$. Dengan mencuplik jarak antar not diharapkan dapat mewakili melodi pada lagu yang diubah menjadi citra pada spektogram.

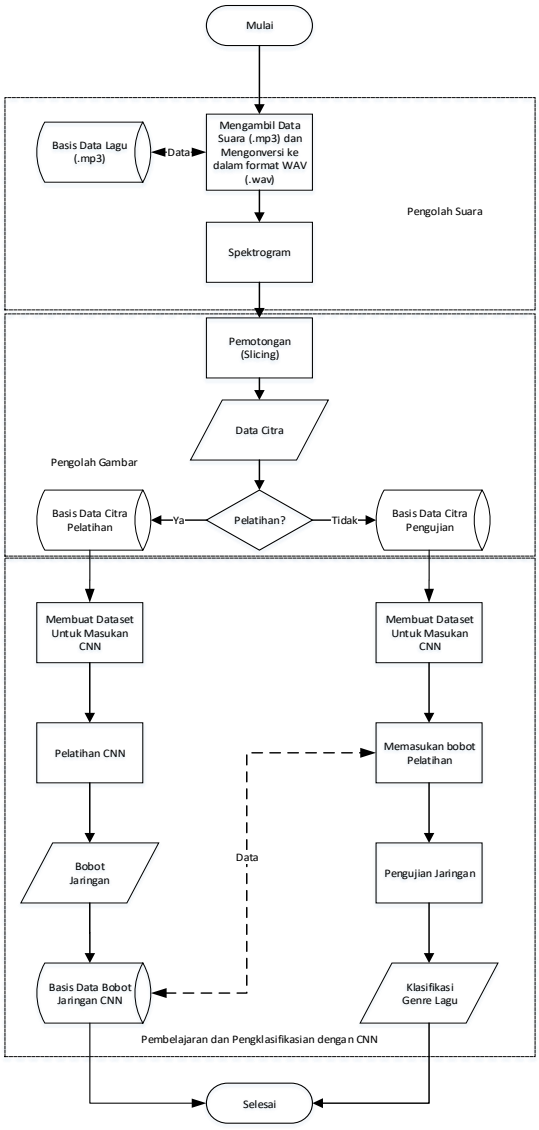
2.3. Pengolah Citra

Pada bagian ini citra spektogram akan dipotong-potong menjadi citra keabuan dengan dimensi yang lebih kecil. Sebuah lagu terdiri dari ketukan yang berulang sepanjang lagu, sehingga untuk mengambil ciri dari sebuah lagu maka hanya diperlukan secuplik informasi saja. Ketukan yang berulang ini memiliki rentang waktu ± 2 detik. Dengan mengasumsikan semua lagu pada basis data memiliki ciri seperti ini, maka untuk mencirikan sebuah genre akan disampling citra sepanjang 2 detik atau 100-pixel. Untuk meminimalisir anomali cuplik ini dan memilih bilangan berpangkat 2 terdekat maka digunakan nilai 128-pixel. Sehingga ciri genre sebuah lagu yang akan dipelajari oleh CNN sebesar 128 × 128 pixel.

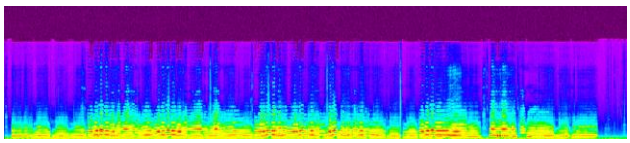
2.4. Convolutional Neural Network

CNN adalah metode perkembangan dari JST yang mana mempelajari pola dengan ransangan neuron pada struktur 3D. CNN dikembangkan untuk menggali informasi dari jumlah data yang lebih besar dibandingkan dengan JST. Pada umumnya proses pembelajaran pada CNN sama dengan proses pada JST yang mana dibagi ke dalam 3 lapisan utama yaitu lapisan masukan, lapisan tersembunyi, dan lapisan luaran. Proses perubahan nilai bobot pada CNN sama dengan JST yaitu dengan menggunakan metode *backpropagation*. Pada tahap pelatihan setiap potongan citra pada basis data latih akan menjadi masukan CNN kemudian luaran pada setiap

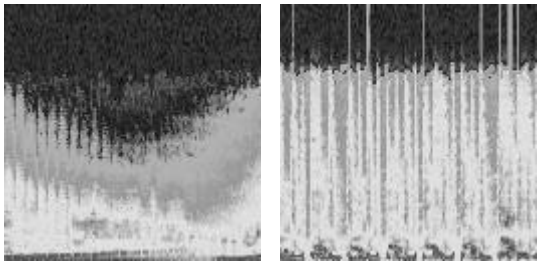
potongan itu akan menghasilkan golongan genre. Galat pada penggolongan genre akan merubah bobot CNN pada setiap iterasi citra. Metode optimasi yang digunakan pada penelitian ini adalah SGD, Adam, Adadelta, dan Adagrad. Pada tahap pengujian, setiap potongan citra pada basis data uji akan menjadi masukan pada CNN kemudian hasil dari penggolongan potongan citra tersebut akan menentukan golongan genrenya.



Gambar 1. Diagram alir sistem penggolongan genre berdasarkan spektogram



Gambar 2. Citra hasil spektogram sepanjang lagu tanpa keterangan (raw).



Gambar 3. Citra keabuan hasil potongan yang berukuran 128x128 pixel

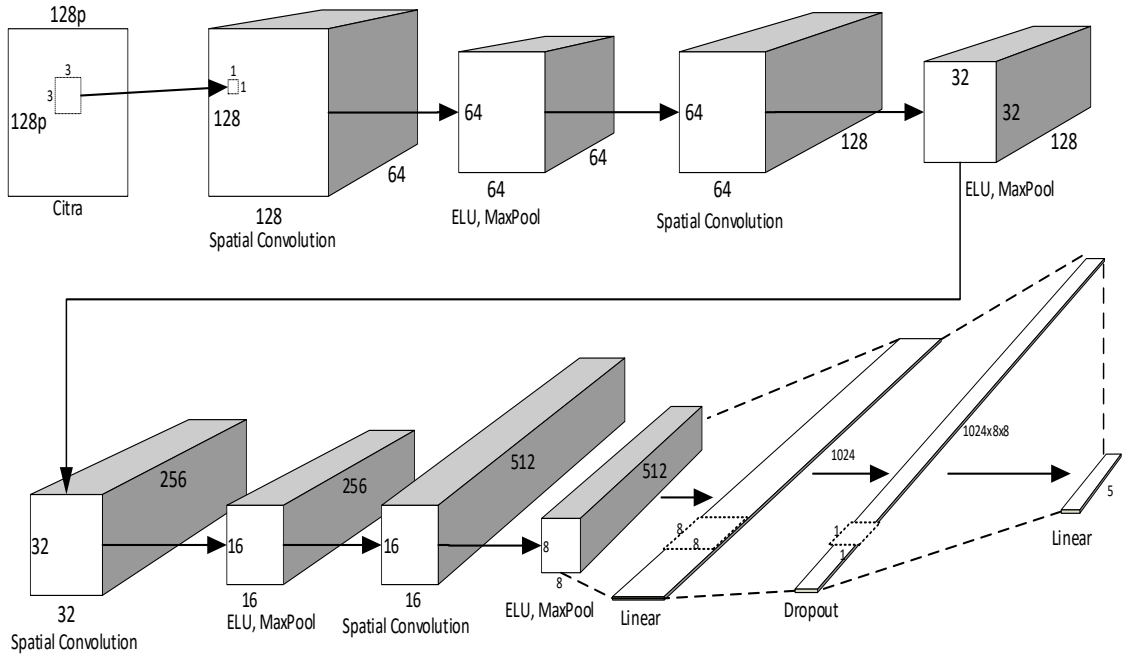
2.5. Model CNN

Struktur CNN dibuat agar CNN dapat mempelajari dan menggolongan sebuah lagu dengan galat sekecil-kecilnya. Pembangunan suatu CNN didasari oleh cara kerja otak manusia dalam menggolongkan suatu benda. Model CNN pada penelitian ini menggunakan metode CNN yang mana suatu citra akan terhubung dengan neuron-neuron 3 dimensi dengan cara kerja konvolusi sepanjang bidang 3 dimensi tersebut. Pada dasarnya suatu informasi berupa sebagian data kecil dari keseluruhan data. Pengenalan informasi dari sebuah data haruslah melaui tahap pengesktrasian ciri. Struktur CNN pada penelitian ini dirancang untuk mengenali informasi dari keseluruhan data dengan memperkecil dimensi citra dengan maxpooling dan pengesktrasian ciri oleh kernel (agen pembelajaran) yang berbeda pada setiap neuron.

Struktur CNN yang dipakai pada penelitian ini adalah dengan menggunakan 18 lapisan yang terdiri dari *Spatial Convolution Layer*, *ELU Layer*, *Spatial Max Pooling Layer*, dan *LogSoftMax Layer*. Struktur CNN yang dipakai pada penelitian ini secara urut sebagai berikut.

1. Spatial Convolution (1,64)
2. ELU
3. Spatial Max Pooling (2,2)
4. Spatial Convolution (64,128)
5. ELU
6. Spatial Max Pooling (2,2)
7. Spatial Convolution (128,256)
8. ELU
9. Spatial Max Pooling (2.2)
10. Spatial Convolution (256,512)
11. ELU
12. Spatial Max Pooling (2,2)
13. View
14. Liniear (512,1024)
15. ELU
16. View
17. Liniear (1024*8*8,5)
18. LogSoftMax

Struktur CNN ini dibangun pada framework Torch7 dengan *library* ‘nn’.



Gambar 4. Struktur CNN

3. Hasil dan Analisa

3.1. Pelatihan dan Pengujian Variasi Filter Pembelajaran

Pada penelitian ini digunakan algoritme pembelajaran SGD dengan variasi filter yaitu 3x3, 5x5, dan 7x7. Penelitian ini akan melihat hasil penggolongan genre berdasarkan variasi filter. Pada Tabel 2 dapat dilihat bahwa perbedaan filter dapat menyebabkan perbedaan proses belajar suatu sistem. Contoh kasus pada lagu "Firepower Records - Hi Im Ghost - Halfway [Your EDM Exclusive Premiere].mp3" yang bergenre dubstep mengalami kesalahan penggolongan pada filter 5×5 dan 7×7. Dari 104 citra potongan pada lagu tersebut, 81 citra tergolong sebagai dubstep dan 9 tergolong reggae pada filter 3×3, 42 tergolong dubstep dan 47 tergolong reggae pada filter 5×5, dan 37 tergolong dubstep dan 61 tergolong reggae pada filter 7×7. Berdasarkan informasi detil pada penggolongan lagu tersebut dapat dilihat bahwa filter 5×5 dan 7×7 memiliki performansi yang kurang baik untuk membedakan genre dubstep dan reggae, sebagai contoh citra potongan pada lagu tersebut yang salah dideteksi dapat dilihat pada Gambar 5.

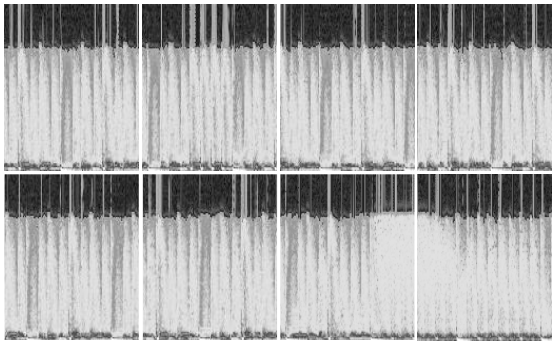
Tabel 2. Perbandingan Hasil Pelatihan dan Pengujian Variasi Filter Pembelajaran

Variasi Filter	Akurasi Pelatihan	Akurasi pengujian
3x3	100%	98%
5x5	100%	82%
7x7	100%	76%

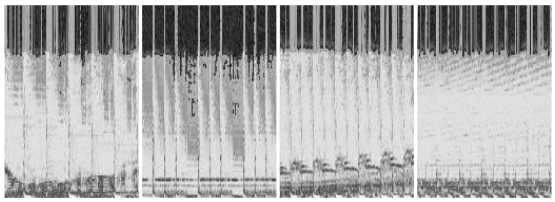
Pada Gambar 4 dapat dilihat bahwa citra uji (a) sekilas memiliki kemiripan pada citra latih (b) dan (c). Namun jika dilihat dengan seksama maka genre dubstep cenderung memiliki pola pada frekuensi tinggi yang berbentuk garis-garis tipis pada citra yang mana reggae tidak memilikinya. Genre dubstep juga cenderung memiliki pola berulang pada tiap frekuensi sedangkan genre reggae cenderung lebih dinamis pada tiap frekuensi. Pada filter 3×3 perbedaan pola ini dapat dipelajari oleh CNN, namun pada filter dengan dimensi yang lebih tinggi yaitu 5×5 dan 7×7 tidak. Ilustrasi perbedaan pembelajaran oleh filter ini ditunjukan pada Gambar 6.

Pada Gambar 5 dapat dilihat perbedaan luas daerah yang dipelajari oleh setiap kernel. Pada filter 3×3 respon pembelajaran cenderung lebih peka terhadap suatu pola rinci dibandingkan dengan filter 5×5 dan 7×7. Citra spektogram yang digunakan sebagai pelatihan memiliki spesifikasi horisontal sebesar 20 ms/bit dan vertikal sebesar 173 Hz/bit sehingga filter 3×3 dapat mengenali pola dengan tingkat kepekaan tiap 60 ms suara dan pergantian satu not. Jika dibandingkan dengan filter 5×5, yang memiliki kepekaan tiap 100ms suara dan 3 kali pergantian not, dan filter 7×7, yang memiliki kepekaan tiap 140 ms suara dan 4 kali pergantian not, filter 3×3 memiliki kepekaan yang lebih baik dalam rentang waktu dan pergantian not (frekuensi).

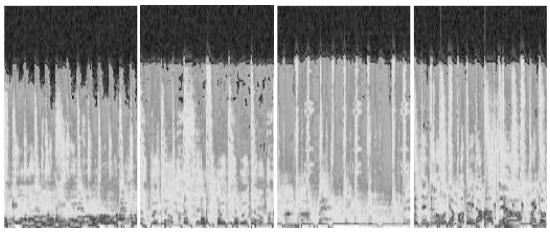
Keunggulan filter 3×3 dengan kepekaannya dalam rentang waktu dan pergantian not menjadikan filter ini cocok untuk mempelajari data latih dengan spesifikasi yang diajukan dalam penelitian ini dengan akurasi 98%. Sedangkan filter lainnya yaitu 5×5 dan 7×7 kurang baik dalam mempelajari data latih dengan spesifikasi yang diajukan dalam penelitian ini dengan masing-masing akurasi sebesar 82% dan 76%.



(a)

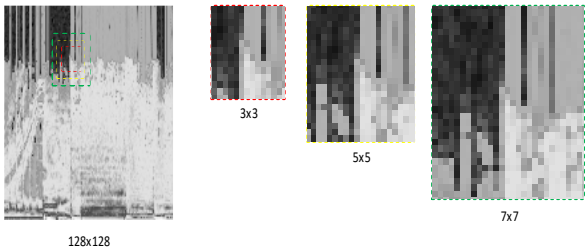


(b)



(c)

Gambar 5. (a) Citra potongan yang salah digolongkan pada filter 5×5 dan 7×7
(b) Citra potongan data latih dubstep
(c) Citra potongan data latih reggae



Gambar 6. Visualisasi Variasi Filter

3.2. Pelatihan dan Pengujian Variasi Algoritme Pembelajaran

Algoritme pembelajaran yang dimaksud adalah algoritme yang dipakai untuk mengubah bobot CNN pada tahap pelatihan. Pada Tabel 3 dapat dilihat bahwa perbedaan algoritme pembelajaran mempengaruhi waktu pelatihan. Faktor yang mempengaruhi kecepatan suatu algoritme adalah formula algoritme tersebut untuk mencapai suatu titik konvergensi. Algoritme SGD memiliki waktu terlama karena hanya menggunakan nilai galat (*error*) dan koefisien pembelajaran sebagai parameter untuk memperbaharui nilai bobot. Algoritme adaptif Adadelta, Adagrad, dan Adam, dilain sisi, memiliki waktu yang cenderung lebih cepat dibandingkan dengan SGD karena memiliki parameter-parameter tambahan untuk memperbaharui nilai bobot.

Tabel 3. Hasil Pelatihan dan Pengujian Variasi Algoritme Pembelajaran

Filter	Waktu Pelatihan
SGD	120 menit
AdaDelta	37 menit
AdaGrad	41 menit
AdaM	35 menit

Dari Tabel 3 dapat dilihat bahwa algoritme SGD butuh waktu yang lebih lama untuk mencapai konvergensi dibandingkan dengan algoritme adaptif lainnya yaitu AdaDelta, AdaGrad, dan Adam. Sedangkan waktu yang dibutuhkan untuk mencapai konvergensi antar ketiga algoritme adaptif relatif sama.

4. Kesimpulasn

Berdasarkan hasil pelatihan dan pengujian sistem penggolongan lagu berdasarkan spektogram dengan CNN yang dilakukan, didapatkan bahwa pelatihan dengan filter 3x3 menggunakan algoritme Adam memiliki performansi terbaik dengan akurasi pelatihan dan pengujian sebesar 100% dan waktu tercepat dibandingkan dengan algoritme adaptif lainnya. Filter 5x5 dan 7x7 memiliki akurasi yang lebih buruk dibandingkan dengan filter 3x3 pada tahap pengujian. Algoritme SGD membutuhkan waktu konvergensi yang lebih lama dibandingan dengan algorimta adatif. Penelitian ini dapat dikembangkan lebih lanjut dengan mengaplikasikannya langsung ke dalam sistem lain yang berhubungan langsung dengan pengguna lagu seperti mesin pencari genre lagu. Penelitian ini juga dapat dikembngkan dengan menggunakan metode analisis suara lainnya seperti *Mel Frequency Ceptral Coefisien, Linea Predictive Coding*, dan lainnya. Penelitian ini juga dapat dikembangkan untuk menggolongkan lebih banyak genre.

Referensi

[1]. A. Human and F. Engineering, “American National Standard,” no. 631, p. 11747, 2013.

[2]. J. (Juyang) Weng, N. Ahuja, and T. S. Huang, “Learning Recognition and Segmentation Using the Cresceptron,” *Int. J. Comput. Vis.*, vol. 25, no. 2, pp. 109–143, 1997.

[3]. Y. M. G. Costa, L. S. Oliveira, and C. N. Silla, “An evaluation of Convolutional Neural Networks for music classification using spectrograms,” *Appl. Soft Comput.*, vol. 52, pp. 28–38, 2017.

[4]. D. Palaz, M. Magimai-Doss, and R. Collobert, "Analysis of CNN-based speech recognition system using raw speech as input," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2015–Janua, pp. 11–15, 2015.

[5]. D. Palaz, M. Magimai.-Doss, and R. Collobert, "Convolutional Neural Networks-Based Continuous Speech Recognition Using Raw Speech Signal," *Icassp*, pp. 4295–4299, 2015.

[6]. R. Indonesia, *Undang-Undang Hak Cipta 1982*, no. 29. 1997.

[7]. E. Taylor, *Music Theory in Practice, Grade 1*. Associated Board of the Royal Schools of Music, 2008.

[8]. D. Speiser, "Discovering the Principles of Mechanics 1600-1800: Essays by David Speiser." Basel/Boston/Berlin, Birkhäuser Verlag AG, hal. 331, 2008.

[9]. R. M.~Bracewell, *The Fourier Transform and Its Applications*, 3rd ed. McGraw-Hill, 1978.

[10]. C. Van Loan, *Computational Frameworks for the Fast Fourier Transform*. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1992.

[11]. J. O. Smith, *Mathematics of the Discrete Fourier Transform (DFT)*. W3K Publishing, 2007.

[12]. J. (Juyang) Weng, N. Ahuja, dan T. S. Huang, "Learning Recognition and Segmentation Using the Cresceptron," *Int. J. Comput. Vis.*, vol. 25, no. 2, hal. 109–143, 1997.

[13]. M. A. Nielsen, *Neural Networks and Deep Learning*. 2015.

[14]. V. Nair dan G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," *Proc. 27th Int. Conf. Mach. Learn.*, no. 3, hal. 807–814, 2010.